Journal of Experimental Education

4780

A QUARTERLY
Edited by
A. S. BARR

With the Cooperation of

ARTHUR T. JERSILD
PALMER O. JOHNSON
H. H. REMMERS
J. WAYNE WRIGHTSTONE

EARCH BUREAU OF EDNL.
517

VOLUME XXIII
September, 1954 - June, 1955

DEMBAR PUBLICATIONS, INC.
303 East Wilson Street
Madison 3, Wisconsin

TABLE OF CONTENTS

Volume XXIII

September, 1954

A Factorial Study of Teaching Ability: Harley Elwood Erickson 1

A Study of Certain Criteria of Teaching Effectiveness: Harold Milton Anderson 41

Temperament and Teaching Success: Harold Wesley Montross 73

December, 1954

Social Competence and Success in Teaching: Arthur Singer, Jr. 99

Teacher Attitudes Toward Curriculum Change: The Effect of Personal Relationships on the Degree of Favorableness: Evelyn I. Banning 133

Dietary Patterns of Negroes in the Community of Scotlandville, Louisiana (Part I)
A Study of the Relationship Between the Diets and the Health of 25 Negro Families in Scotlandville, Louisiana (Part II): P. E. Thrift and R. B. Ward 157

Reversals, Reversals, Reversals: Helen Kennedy 161

The Image of the Instructor as it is Related to Class Size: John B. Holland 171

Some Educational Implications of the Influence of Rejection on Aspiration Levels of Minority Group Children: Regina M. Goff 179

March, 1955

Statistical Analysis of Scores from Counterbalanced Tests: Julian C. Stanley 187

The Influence of Socio-Cultural Characteristics of Educational Opportunities in Public School Instrumental Music: Melvin L. Zack 209

Simultaneous Absolute Scaling for Several Groups: W. A. Gibson 221

An Experimental Evaluation of Two Methods of Teaching Music in the Fourth and Fifth Grades: Carl B. Nelson 231

A Comparison of Verbal and Pictorial Self-Rating-Scale Categories: Julian C. Stanley 239

The Interpretation of the Standard Error of Measurement: C. H. Patterson 247

Simultaneous Examination and Method Analysis by Variance Algebra: William J. Moonan 253

Classification by Multivariate Analysis with Objectives of Minimizing Risk, Minimizing Maximum Risk, and Minimizing Probability of Misclassification: M. Clemons Johnson..... 259

The Poisson Distribution in Educational Research: Sam Duker........................ 265

Forecasting Future Enrollment by Curve-Fitting Techniques: H. P. Kuang................ 271

An Application of Multiple Regression Analysis in Determining the Relative Contribution of Certain Components of Reading Ability in Grade and High School Achievement: Lavern L. Krantz 275

A Reduced Formula for Point-Biserial Coefficient of Correlation: William O. Buschman..... 279

June, 1955

Prediction of the Probability of Graduation of Engineering Transfer Students: J. Stanley Ahmann 281

Characteristics of the Effective Teacher Based on Pupil Evaluations: Percival M. Symonds... 289

Significant Characteristics of College Students with Varying Flexibility in Reading Rate:
  I. Eye-Movements in Reading Prose
  II. Motor and Perceptual Skill in "Reading" Material Whose Meaning is Unimportant: Frank Laycock 311

An Evaluation Instrument for the Analysis of Teacher Effectiveness: Granville B. Johnson... 331

Attitudes and Professional Relationships of Teachers: A Study of Teacher Morale: E. C. Hunter 345

The Influence of a Child Purpose Upon the Perseverance of Young Children: Louis Ada Wilson 353

What is an I. Q. ? Florence Brumbaugh 359

Partial Reinforcement Related to "Free" Responding in Extinction with Pre-School Children: Nicholas A. Fattu, Edmund V. Mech, Donavon Auble.......................... 365

An Appraisal of Teacher-Counselor Adviser Teaching: Russell L. Jenkins 369

Peer Group Rating of High School Pupils with High and Low School Marks: Evan R. Keislar .. 375

Regression Coefficients vs. Correlation Coefficients: John Schmid, Jr..................... 379

Journal of Experimental Education

Volume XXIII September, 1954 Number 1

A FACTORIAL STUDY OF TEACHING ABILITY

HARLEY ELLWOOD ERICKSON *
Superior State College
Superior, Wisconsin

478U

SECTION I

INTRODUCTION

MANY STUDIES of a followup nature have been made of graduates of teacher training curricula to determine the success of these persons as teachers. Such studies have been, in one manner or another, concerned with the assessment and measurement of teaching ability. That such evaluation is necessary has been pointed out many times. Barr[1] says "the measurement of teaching ability is important: (1) in the training of teachers in service; (2) in the institutional training of teachers; (3) in the administration of the teaching staff; and (4) in research relating to these fields." Teacher training institutions are confronted with the dual problem of recruiting and selecting those candidates who seem to show the greatest promise of becoming successful teachers and of providing those curricular and extra-curricular activities which will aid these candidates when they have entered the teaching profession. Evaluation is important in both.

Recent years have seen the role of the teacher undergo significant changes. Baxter[2] points out: "It is no longer enough that the teacher be the possessor of knowledge. Today's teacher must be a 'social engineer', capable of setting up a provocative environment for children's learning, charting the course of each individual child through the ever-changing social relationships in which he is involved and assisting each pupil to grow in his understanding of himself and of others." One can readily realize the implications of statements such as this for the recruitment, selection, training, and placement of teachers.

Many persons have attempted to construct instruments of one sort or another that might be used in teacher evaluation. As yet, educators have not, however, reached agreement concerning what constitutes an adequate criterion of teaching success. In general, research studies have employed criteria of the following types: (1) a rating assigned by the principal or superintendent; (2) a rating by the teacher's supervisor or a visiting supervisor; (3) an evaluation of the teacher by the pupils; and (4) an evaluation by the teacher's peers. Less frequently a criterion of pupil gain has been used, but when this has been utilized the gains have been expressed as difference scores obtained by subtracting a final test score from a pre-test score. Most educators would agree that the teacher should provide an atmosphere which will foster desirable growth in students, but they would not all agree that this growth is to be assessed only by achievement tests in subject matter areas. On occasion a self-evaluation has been advocated and used. Often a composite criterion has been formed by combining several of these criteria.

Another problem confronting researchers is that of determining those traits or characteristics which appear to be related to teaching success. That many of these traits probably lie outside the area of subject matter competence has been pointed out by Olson[3] in his comments on various studies made of teachers: "It is interesting and significant in the various studies that knowledge of the subject matter as such seems to be taken for granted and that the success or failure of a teacher appears to lie more largely in personal qualities and social relation-

---

* The author has incurred many obligations in carrying out this study, in particular to Dr. A. S. Barr, his major professor, for his constant encouragment, assistance, and leadership throughout the course of this investigation.

1. Arvil S. Barr. "The Measurement of Teaching Ability," Journal of Educational Research, XXVIII (April 1935), pp. 561.

2. B. Baxter. Teacher Pupil Relationships (New York: Macmillan Co., 1950).

3. Willard C. Olson. Child Development (Boston: D. C. Heath & Co., 1949), p. 285.

ships."

That such qualities might be related to the temperament or personality of the teachers was pointed out in a UNESCO pamphlet:[4] "Various research studies show clearly that the emotional stability of teachers affects that of pupils. What happens to boys and girls in school depends in large measure on the personal growth and development of the teachers with whom they have to work. Unhappy, frustrated, dissatisfied teachers cannot help their pupils to become happy, well-adjusted young people."

Statement of the Problem

This dissertation is concerned with the relationships between supervisory estimates of teaching success and various measures of teacher achievement, temperament, and personality.

Specifically this study is concerned with developing equations for the prediction of teaching success with a minimum number of variables and maximum stability as suggested by factor analysis of a variety of measures of teacher characteristics and efficiency. In order to develop a basis for simplified patterns of prediction a factorial study is made of the interrelationships among variables. This is done to determine factors common to estimates of teaching success and data-gathering devices commonly employed to measure the achievement, temperament, and personality of teachers.

In brief, this study will attempt to answer the following questions:

1. Will a general factor account for the matrix of intercorrelations of the nine estimates of teaching success?
2. How many common factors are necessary to account for the intercorrelations of the nine estimates of teaching success?
3. Will estimates of these factors in the form of factor scores serve as a basis for separating successful teachers from unsuccessful teachers?
4. Will a grouping of variables and an analysis similar to that used in (2) above, applied to three rectangular matrixes of orders seven by nine, sixteen by nine, and ten by nine, respectively, account for the correlations in these matrices?
5. What measures of achievement, temperament, and personality are significantly related to these factor scores used as estimates of teaching success?

SECTION II

DESIGN OF THE STUDY

Description of the Subjects

SIXTY-FOUR teachers engaged in their second year of teaching in Wisconsin high schools were chosen for this study. These sixty-four cases are a sub-group of the total group of 101 beginning teachers visited during the beginning teacher visitation program conducted the latter half of the first semester of the 1950-1951 school year. Those visited had graduated from the University of Wisconsin in 1950 with the University Teachers Certificate and had obtained their first teaching positions in the State of Wisconsin in the fall of the year 1950. The thirty-seven teachers not included for study have left the teaching profession: some married, others became members of the armed forces, and several went into other vocational areas. The sixty-four remaining cases consist of thirty-four male and thirty female teachers.

These teachers represent nineteen fields of major preparation. A tabulation of the majors and minors will be found in Table I. Twenty-eight subject areas are represented for which these teachers meet University of Wisconsin major or minor requirements or State of Wisconsin minor field certification requirements. The total frequency of majors is 67 as contrasted with the 64 teachers. This apparent discrepancy arises since teachers numbered 15 and 46 had double majors of History and Economics, and teacher No. 47 had a double major of Speech and English.

The mean total grade point average for the group of sixty-four teachers studied was 1.94 with a standard deviation of .40. This compares very well with the mean of 1.88 and standard deviation of .38 calculated for 205 (1949) school of education graduates and holders of the University Teachers Certificate. This group does not appear markedly different from the 1949 graduates in terms of means and standard deviations. It cannot be determined whether this comparison would hold over a period of years, since statistics are not available. However, Schmid[5] in a study of 102 (1948) graduates reported a mean of 1.75 and a standard deviation of .40 for total grade point average.

The mean Total Score Percentile Rank on the

---

[4] U. N. E. S. C. O. The Education and Training of Teachers Toward World Understanding.

[5] John Schmid, Jr. "Factor Analyses of Prospective Teachers' Differences," Journal of Experimental Education, XVIII (June 1950), pp. 287-319.

TABLE I

TABULATION OF MAJORS AND MINORS FOR THE 64 TEACHERS

	University Majors			University Minors			State Minors		
Subject Area	Male	Female	Total	Male	Female	Total	Male	Female	Total
English	5	8	13	1	1	2	15	9	24
History	6	1	7	1	2	3	5	—	5
American Instit.	3	—	3	—	—	—	—	—	—
Agriculture	1	—	1	—	—	—	—	—	—
Physical Educ.	11	—	11	1	—	1	—	—	—
Home Economics	—	7	7	—	—	—	—	—	—
Business Educ.	1	3	4	—	—	—	—	—	—
Physics	1	—	1	—	—	—	—	—	—
Chemistry	1	—	1	—	6	6	—	1	1
Natural Science	—	3	3	—	—	—	—	—	—
French	—	2	2	1	1	2	—	1	1
Spanish	—	1	1	—	5	5	—	3	3
Speech	1	1	2	—	1	1	1	—	1
Speech Correction	—	2	2	—	—	—	—	—	—
Economics	2	—	2	5	1	6	1	3	4
Art Education	3	1	4	—	—	—	—	—	—
Sociology	1	—	1	4	2	6	—	1	1
Music	1	—	1	—	—	—	—	—	—
Recreation	—	1	1	—	—	—	—	—	—
General Science	—	—	—	12	1	13	—	—	—
Political Science	—	—	—	1	—	1	—	—	—
Mathematics	—	—	—	3	—	3	1	—	1
Botany	—	—	—	—	1	1	—	—	—
Biology	—	—	—	1	—	1	—	—	—
German	—	—	—	—	1	1	—	1	1
Geography	—	—	—	—	—	—	1	1	2
Zoology	—	—	—	—	—	—	—	2	2
Italian	—	—	—	—	—	—	1	—	1
Total	37	30	67	30	22	52	25	22	47

American Council on Education Psychological Examination is 60.3. Bach[6] in a study involving seventy-six 1950 graduates obtained a mean of 65.8 while Lamke[7] reported a mean ACE percentile score of 72 for thirty-two 1950 graduates.

The mean percentile rank in high school class is 75.7.

It is thought that this group is reasonably representative of recent School of Education graduates at the University of Wisconsin.

The Data-Gathering Devices Employed

Forty-two variables were used in the analysis which follows. Nine of the variables are estimates of teaching success; seven were derived from the Thurstone Temperament Schedule; sixteen from the 16 P.F. Test; and the remaining ten from University records.

The forty-two variables are listed below. The number preceding each measure is used to identify that variable in subsequent correlation and factor matrices. A code symbol follows each measure.

1.	Acceptability rating by principal	PA
2.	Principal's M-blank rating (1st year)	PM1
3.	Supervisor I M-blank rating	S1M
4.	Supervisor II M-blank rating	S2M
5.	Principal's M-blank rating (2nd year)	PM2
6.	Teacher rating by an outside agency on a special rating form	SPR
7.	Teacher's self evaluation	SE
8.	Peer evaluation	PE
9.	Pupil evaluation	PuE
10.	Active (Thurstone)	A-T
11.	Vigorous (Thurstone)	V-T
12.	Impulsive (Thurstone)	I-T
13.	Dominant (Thurstone)	D-T
14.	Stable (E for emotionally stable) (Thurstone)	E-T
15.	Sociable (Thurstone)	S-T
16.	Reflective (Thurstone)	R-T
17.	Cyclothymia vs. Schizothymia (16 P.F. Test)	A
18.	General Intelligence vs. Mental Defect (16 P.F. Test)	B
19.	Emotional Stability or Ego Strength vs. General Neuroticism (16 P.F. Test)	C
20.	Dominance or Ascendance vs. Submission (16 P.F. Test)	E
21.	Surgency vs. Desurgency (16 P.F. Test)	F
22.	Positive Character vs. Immature Dependent Char. (16 P.F. Test)	G
23.	Adventurous Cyclothymia vs. Inherent Withdrawn Schizothymia (16 P.F. Test)	H
24.	Emotional Sensitivity vs. Tough Maturity (16 P.F. Test)	I
25.	Paranoid Schizothymia vs. Trustful Accessibility (16 P.F. Test)	L
26.	Bohemianism vs. Practical Concernedness (16 P.F. Test)	M
27.	Sophistication vs. Rough Simplicity (16 P.F. Test)	N
28.	Worrying Suspiciousness vs. Calm Trustfulness (16 P.F. Test)	O
29.	Radicalism vs. Conservatism (16 P.F. Test)	Q_1
30.	Independent Self-Sufficiency vs. Lack of Resolution (16 P.F. Test)	Q_2
31.	Will Control and Character Stability (16 P.F. Test)	Q_3
32.	Nervous Tension (16 P.F. Test)	Q_4
33.	Practice Teaching grade in Education 75	PT75
34.	Practice Teaching grade in Education Methods Course	PTEM
35.	Grade in Education 73	G73
36.	Grade in Education 74	G74
37.	Lecture Section Grade, Education 75	LG75
38.	Percentile Rank in High School Class	PRHS
39.	American Council on Education Psychological Examination Percentile Rank	ACEPR
40.	Freshman-Sophomore Grade Point Average	Fr.-So. GPA
41.	Junior-Senior Grade Point Average	Jr.-Sr. GPA
42.	Composite: 2 credit major Educational Methods lecture section grade and 2 credit minor methods grade	EM4

Sources of Data

The following data were obtained from the beginning teacher visitation folders:

1. Principal's Acceptability rating.
2. Principal's rating on the Wisconsin Adaptation of the M-Blank.
3, 4. Two visiting supervisor's ratings on the M-Blank.

The following estimates of teaching success were obtained during the 1951-1952 school year:

---

[6] Jacob O. Bach. Practice Teaching Success in Relation to Other Measures of Teaching Ability, Ph.D. Dissertation, University of Wisconsin, 1951.

[7] Tom A. Lamke. "Personality and Teaching Success," Journal of Experimental Education, XX (December 1951), pp. 217-259.

5. Principal's evaluation using the Wisconsin Adaptation of the M-Blank.

6. Rating of the teacher by an outside agency on a special rating form.

7. Self-evaluation.

8. Peer evaluation.

9. Pupil evaluation.

Data for assessing temperament traits were obtained by administering the Thurstone Temperament Schedule to each subject. (Variables 10 through 16.)

Data for assessing the sixteen source traits of personality were obtained by administering the 16 Personality Factor Test to each teacher. (Variables 17 through 32.)

The following data were obtained from University records:

33. Practice Teaching grade in Education 75.

34. Practice Teaching grade in Education Methods Course.

35, 36, and 37. Grades in Education 73, Education 74, and Lecture Section grade in Education 75.

38. Percentile rank in high school class.

39. American Council on Education Psychological Examination Percentile Rank.

40. Freshman-Sophomore Grade Point Average.

41. Junior-Senior Grade Point Average.

42. Lecture Section grade in major methods course and grade in 2 credit minor methods course.

Description of the Measures Used

1. Principal's acceptability rating, PA—This criterion was cooperatively defined by Lamke[8], Ringness[9], and Bach[10] to be that quality of the teacher which leads other people to feel that she should be retained in her present position or possibly promoted to a more responsible assignment. In the fall of 1950, teams of visiting supervisors from the University of Wisconsin visited each beginning teacher. One graduate student from a core group of six graduate students was usually included in each team of two supervisors which visited each teacher during the same class period. This graduate student conducted the interview with the principal to obtain the acceptability rating. The following typical questions were asked at the end of the interview:

1. How do you like this teacher?
2. Does she get along well with the students?
3. Do the parents feel the same way as the pupils do about the teacher?
4. Is the teacher happy here? Does she have friends on the faculty, in the community, etc?
5. What strengths or weaknesses does this teacher have?
6. In terms of beginning teachers with whom you have worked in the past and in light of our discussion as a whole, would you consider this teacher as (1) among the best, (2) above average, (3) average, (4) below average, or (5) failing as a beginning teacher?

The principal was thus led to consider some of the teacher's more important relationships before he was asked for his acceptability rating which is his response to question (6). A low score indicates a high rating.

2. Principal's M-Blank rating (1st year), PM1—A two-page form of the Wisconsin Adaptation of the M-Blank was used during the beginning teacher visitation program (see Appendix B)*. The principal was asked for his rating of each teacher using this form. The "General Evaluation" rating was the score used in this study. The categories on a five-point scale were as follows: (1) outstanding, (2) above average, (3) average, (4) below average, and (5) poor. A low score indicates a high rating.

3. Supervisor I M-Blank rating, S1M, and

4. Supervisor II M-Blank rating, S2M—As indicated previously, one graduate student of the core group was usually included on each visitation team. The rating assigned by this student was classified as Supervisor I rating and the rating of the other visiting supervisor, Supervisor II. In several cases, two students of this core group formed the team. In these instances the Supervisor I and Supervisor II ratings were classified alphabetically. In the case of each of these ratings a low score indi-

---

[8] Ibid.

[9] Tom A. Ringness. Relationships Between Certain Attitudes Towards Teaching and Teaching Success, Ph.D. Dissertation, University of Wisconsin, 1951.

[10] Op. cit.

* All references to Appendices may be found in original thesis on file in Library, University of Wisconsin.

cates a high rating.

5. Principal's M-Blank rating (2nd year), PM2—A form letter and a one-page form of the Wisconsin Adaptation of the M-Blank (see Appendix C) was sent to each principal in January of this year. The rating under the item "General Evaluation" was the score used in this study. A low score indicates a high rating.

6. Teacher rating by an outside agency on a special rating form, SPR—A low score indicates a high rating.

7. Teacher's self-evaluation, SE—Harold M. Anderson of the University of Colorado adapted a pupil evaluation form develped by Bryan for use by pupils, peers, and the teacher. The Teacher's Self Evaluation form (see Appendix D) was sent to each teacher during the past school year. The overall evaluation (item 10) was the score used in this study. The categories of item 10 were as follows: (1) best, (2) second, (3) middle, (4) fourth, and (5) lowest. A low score indicates a high rating.

8. Peer evaluation, PE—Two fellow teachers were sent a form letter asking them to rate the teacher using the Peer Evaluation of Instruction Form (see Appendix E). An attempt was made to have each teacher rated by one male and one female colleague. Those asked to rate the teacher were chosen, using the Official Wisconsin School Directory, from the same subject area, insofar as possible, as the teacher to be rated. The response to item 9 with the following categories, upper, second, middle, fourth, and lowest, was the rating used in this study. The sum of these two peer ratings was the peer evaluation score used. A low score indicates a high rating.

9. Pupil evaluation, PuE—Each teacher's principal was sent a form letter and five Pupil Reaction to Instruction forms (see Appendix F). He was asked to secure the reactions of five students chosen at random from one of the classes taught by the teacher. Each student returned his completed form in a self-addressed envelope furnished by the Teacher Personnel Research Committee. The sum of these five ratings on the general category rating with (1) best fifth, (2) second fifth, (3) middle fifth, (4) fourth fifth, and (5) poorest fifth was the pupil evaluation rating score used in this study. A low number indicates a high rating.

Before passing on to the description of other data-gathering devices it might be well to discuss certain terms.

What is teaching success?

Educators agree that teaching is an exacting and complex act. Certain components such as the unique physical setting of the teaching situation, and the presence of one or more persons in this setting are prerequisites to the teaching act. Each individual, including the teacher, brings to such a setting his own unique pattern of behavior. No two teaching situations are exactly alike due to the complex and intricate interplay of these unique components. By and large, a teacher in a given situation will be judged successful or unsuccessful, effective or ineffective, in terms of how he adjusts his own unique pattern of behavior to the unique physical setting and to the unique behavioral patterns of those with whom he has contacted.

The terms "successful" and "effective" as applied to teachers need not be and probably are not synomous. If by an effective teacher we mean one who provides an atmosphere which will foster desirable growth in pupils, then agreement would be needed on what desirable pupil changes we wish to measure and determine means of assessing these. Barr[11] in 1935 stated: ".... the ultimate criterion of teaching success will have to be found in the changes produced in pupils measured in terms of the objectives of education.... These changes or products of instruction will have to be considered broadly. We shall have to take into consideration not merely knowledges and skills and the more tangible outcomes of instruction but attitudes and ideals and the less tangible outcomes of instruction." If this could be done the degree of effectiveness of a teacher could be determined according to prescribed standards. Up to the present time this has not been adequately accomplished for any grade level or for any subject matter area. Effectiveness so determined could be considered either as an index of success or as an index of effectiveness for that teacher.

Success is a broader term that not only involves the teacher as a director of learning but also involves the teacher as a friend and counselor of pupils, as a member of a professional staff, and as a citizen in a community. To be successful in this broader sense, a teacher not only must produce desirable changes in pupils but he must be acceptable to the administration, to his peers, to the pupils, to the parents, and to the community. Many people will feel that to be effective (as defined in this study), a teacher must possess these broader competencies.

The evaluations and judgments made about teachers by principals and supervisors continue to be important in education. Very often these evaluations are given in the form of ratings such as those used in this study. On the basis of such ratings teachers often are promoted, are given pay increases, and are granted licenses and certificates. Such a rating is the rater's opinion of the teacher expressed as a value judgment. The question of reliability does not arise if we view this rating as the rater's judgment of this teacher's effectiveness relative to the time and to the specific situation in which she was observed. The question of reliability may arise, however, if at some later date the same rater is asked to rate this same teacher again using the same rating instrument. If one asks the question: "Do the two ratings agree?" a question of reliability clearly arises. If the two ratings do not correlate highly we cannot be sure whether this lack of agreement is due to a change in the rater, a change in the ratee, or a change in the interpretation or application of the instrument.

---

[11] Arvil S. Barr. "The Measurement of Teaching Ability," Journal of Educational Research, XXVIII (April 1935), pp. 568.

Each of the ratings used in this study were obtained as judgments of the principal, a visiting supervisor, the teacher's peers, and the pupils taught by the teacher concerning the effectiveness of a particular teacher at a particular time in a particular situation. The specific instruments utilized to obtain these judgments have been described in detail earlier in this section. In this study we will study the interrelationships found to exist between these several rating judgments to determine if these judgments are measuring different factors. Described next are certain measures of temperament.

10. through 16. Thurstone Temperament Schedule, TTS—This questionnaire was developed to show types of temperament. Seven areas of temperament are appraised by questions pertaining to likes and dislikes, preferences and habits, in everyday life. The schedule contains 140 such questions. The twenty items scored for each area are scattered throughout the schedule. The testee answers each question with Yes, No, or ?. However, only the Yes or No responses score on a self-scoring answer pad accompanying the booklet. A copy of the Thurstone Temperament Schedule and the mimeographed letter of instructions sent to each teacher are included in Appendix G.

Test-retest reliabilities for these seven temperament areas were computed for 81 male executives. The retests were given within six months of the first administration. The reliabilities were: A, 0.78; V, 0.78; I, 0.79; D, 0.82; E, 0.61; S, 0.73; and R, 0.75.

Split-half reliabilities corrected by the Spearman Brown formula for 200 college men were: A, 0.48; V, 0.61; I, 0.65; D, 0.77; E, 0.63; S, 0.68; R, 0.73. Split-half reliabilities computed for 157 college women in a similar manner were: A, 0.46; V, o.63; I, 0.65; D, 0.77; E, 0.64; S, 0.73; and R, 0.62.

The seven areas of temperament assessed by this schedule are described below. These descriptions are quoted from the Examiner Manual for the Thurstone Temperament Schedule.

10. Active, A-T

A person scoring high in this area usually works and moves rapidly. He is restless whenever he has to be quiet. He likes to be "on the go" and tends to hurry. He usually speaks, walks, writes, drives, and works rapidly, even when these activities do not demand speed.

11. Vigorous, V-T

A person with a high score in this area participates in physical sports, work requiring the use of his hands and the use of tools, and outdoor occupations. The area emphasizes physical activity using large muscle groups and great expenditure of energy. This trait is often described as "masculine", but many women and girls will score high in this area.

12. Impulsive, I-T

High scores in this category indicate a happy-go-lucky, daredevil, carefree, acting-on-the-spur-of-the-moment disposition. The person makes decisions quickly, enjoys competition, and changes easily from one task to another. The decision to act or change is quick regardless of whether the person moves slowly or rapidly (Active), or enjoys or dislikes strenuous projects (Vigorous). A person who doggedly "hangs on" when acting or thinking is typically low in this area.

13. Dominant, D-T

People scoring high on this factor think of themselves as leaders, capable of taking initiative and responsibility. They are not domineering, even though they have leadership ability. They enjoy public speaking, organizing social activities, promoting new projects, and persuading others. They are the ones who would probably take charge of the situation in case of an accident.

14. Stable (E for emotionally stable), E-T

Persons who have high scores usually are cheerful and have an even disposition. They can relax in a noisy room, and they remain calm in a crisis. They claim that they can disregard distractions while studying. They are not irritated if interrupted when concentrating, and they do not fret about daily chores. They are not annoyed by leaving a task unfinished or by having to finish it by a deadline.

15. Sociable, S-T

Persons with high scores in this area enjoy the company of others, make friends easily, and are sympathetic, cooperative, and agreeable in their relations with people. Strangers readily tell them about personal troubles.

16. Reflective, R-T

High scores in this area indicate that a person likes meditative and reflective thinking and enjoys dealing with theoretical rather than practical problems. Self-examination is characteristic of reflective persons. These people are usually quiet, work alone, and enjoy work that requires accuracy and fine detail. They often take on more than they can finish, and they would rather plan a job than carry it out.

The following consideration regarding the use of the term temperament was thought important for this study.

Allport[12] speaks of temperament as being a certain class of raw material from which personality is fashioned. "Strictly speaking there is no temperament apart from personality, nor any personality devoid of temperament." He makes the following definition of temperament:[13] "Temperament refers to the characteristic phenomena of an individual's emotional nature, including his susceptibility to emotional stimulations, his customary strength and speed of response, the quality of his prevailing mood, and all peculiarities of fluctuation and intensity in mood; these phenomena being regarded as dependent upon constitutional make-up, and therefore largely hereditary in origin." Those areas of temperament assessed by the Thurstone Temperament Schedule will be considered in this study. Seven areas of temperament are appraised by this questionnaire. Thurstone[14] states that his schedule was ". . . . designed to assess those traits which are relatively permanent for each person. . . . It is limited to a practical description of important aspects of temperament."

17. through 32. 16 P. F. Test (Cattell), 16 PF

This self report questionnaire was constructed to assess the sixteen source traits which Cattell has postulated account for the fundamental structure of personality. The validity of the test is logical. The test consists of two forms, A and B. Each form contains 187 questions each of which may be answered by possible responses of Yes, ?, No. Cattell recommends the forced choice response with the elimination of "?" under certain conditions. This recommendation was followed in this study. The questions which score for each factor are scattered throughout the test. A response to a given question contributes to the score of only one factor.

The split-half reliability of the test on a sample of 200 of the general population, corrected to the full number of items in the A and B forms, is reported in the test manual to be as follows for the sixteen factors: A, 0.84; B, 0.70; C, 0.71; E, 0.82; F, 0.85; G, 0.56; H, 0.74; I, 0.54; L, 0.55; M, 0.72; N 0.65; O, 0.88; Q_1, 0.50; Q_2, 0.61; Q_3, 0.53; Q_4, 0.76.

Both forms were administered to each subject. A copy of the 16 P. F. Test and the mimeographed letter of instructions sent to each teacher is included in Appendix H.

The sixteen source traits of personality assessed by the test are described below primarily in terms of questionnaire responses. The plus pole corresponds to the category listed on the left and is denoted by a high numerical score for that factor. For more detailed descriptions of these factors the reader is referred to many of Cattell's written reports (see Bibliography) or Guide to Principles and Techniques of Personality and Ability Testing, and Handbook for the Sixteen Personality Factor Questionnaire. The following partial descriptions of these factors are quoted from the latter publication:

17. Factor A. Cyclothymia vs. Schizothymia, A

A score of 40 is possible for this factor. This factor has been found to load most highly the following traits:

Good natured, easy going vs. spiteful, grasping, critical
Ready to cooperate vs. obstructive
Attentive to people vs. cool, aloof
Soft hearted, kindly vs. hard
Trustful vs. suspicious
Adaptable vs. rigid
Warm hearted vs. cold

. . . . In questionnaire responses the A+ individual expresses marked preference for occupations dealing with people and socially impressive situations while the A- person likes things or words (particularly material things), working alone, intellectual companionship, and avoidance of clash of viewpoints.

---

[12] Gordon W. Allport. Personality, A Psychological Interpretation (New York: Henry Holt & Co., 1937), p. 63.

[13] Ibid., p. 54.

[14] L. L. Thurstone. Examiner Manual for the Thurstone Temperament Schedule (Chicago: Science Research Associates, Inc., 1950), p. 1

18. Factor B. General Intelligence vs. Mental Defect, B

A score of 26 is possible for this factor. The measurement of intelligence carries with it as a factor in the personality realm some of the following ratings:

Conscientious vs. somewhat unscrupulous
Persevering vs. quitting
Intellectual, cultured vs. boorish

....and indicating some moderate tendency for the more intelligent person to have somewhat more of the traits which we often call character traits.

19. Factor C. Emotional Stability or Ego Strength vs. General Neuroticism, C

A score of 52 is the highest possible score for this factor. This factor loads:

Emotionally mature vs. lacking in frustration tolerance
Emotionally stable vs. changeable
Calm, phlegmatic vs. showing general emotionality
Realistic about life vs. evasive
Absence of neurotic fatigue vs. neurotically fatigued
Placid vs. worrying

This factor is one of good versus poor dynamic integration of the personality.... In the questionnaire manifestation the C- person is easily annoyed by things and people, is dissatisfied with the world situation, his family, the restrictions of life and his own health.

20. Factor E. Dominance or Ascendance vs. submission, E.

A score of 52 is possible for this factor.
Assertive, self-assured vs. submissive
Independent minded vs. dependent
Hard, stern vs. kindly, soft-hearted
Solemn vs. expressive
Unconventional vs. conventional
Tough vs. easily upset
Attention getting vs. self-sufficient

....At present no good objective tests exist for this factor, since it seems to show itself more in human relations than in assertion over inanimate difficulties. The chief questionnaire responses indicate a tendency to assert oneself socially, to believe that one is right in disagreements, to borrow or trespass without embarrassment, and to be ready to take a chance in dangerous situations.

21. Factor F. Surgency vs. Desurgency, F

A score of 52 is possible for this factor. This factor loads:

Talkative vs. silent, introspective
Cheerful vs. depressed
Placid vs. anxious
Frank, expressive vs. incommunicative, smug
Quick, alert vs. languid, slow

....The questionnaire responses show F+ to be carefree, happy-go-lucky, fond of bustle and excitement, inclined to practical jokes and disinclined to occupations requiring close and accurate work. There is a significant trend from F+ behavior with age.

22. Factor G. Positive Character vs. Immature Dependent Character, G

A score of 40 is the highest possible score for this factor.

Persevering, determined vs. quitting, fickle
Responsible vs. frivolous
Emotionally mature vs. demanding, impatient
Consistently ordered vs. relaxed, indolent
Conscientious vs. undependable
Attentive to people vs. obstructive

....Subjectively the G+ person views himself as correct in, and a guardian of manners and morals, persevering, planful, able to concentrate, interested in analyzing people, cautious and preferring efficient people to other companions.

23. Factor H. Adventurous Cyclothymia vs. Inherent Withdrawn Schizothymia, H

A score of 52 is possible for this factor. This factor loads:

Gregarious sociability vs. shyness, withdrawing tendency
Adventerous, bold vs. cautious, retiring
Having marked interest in the opposite sex vs. slight interest in the opposite sex
Frivolous vs. conscientious
Strong, artistic, or sentimental interests vs. lack of same
Abundant emotional response vs. coolness, aloofness

....The H- individual reports himself to be intensely shy, convinced of his inferiority, slow and impeded in expressing himself, disliking occupations with personal contacts, preferring one or two close friends to large groups and not able to keep in contact with all that is going on around him.

24. Factor I. Emotional Sensitivity vs. Tough Maturity, I

A score of 40 is possible for this factor. This factor loads:

Demanding, impatient vs. emotionally mature
Dependent, immature vs. independent minded
Imaginative, introspective vs. set and smug
Kindly, gentle vs. hard, cynical
Aesthetically fastidious vs. lacking artistic feeling
Frivolous vs. responsible
Attention getting vs. self-sufficient

....In the questionnaire realm, such a factor has appeared several times and has been called self-sufficiency, but as there are two self-sufficiency factors, this needs special description, particularly since attention-getting appears in the alleged self-sufficiency. The individual typically shows a fastidious aversion to "crude" people and occupations but a liking for travel and new experiences along with a labile, imaginative, aesthetic mind and a certain impracticality in general affairs.

25. Factor L. Paranoid Schizothymia vs. Trustful Accessibility, L

A score of 40 is possible for this factor. This factor loads:

Prone to jealousy vs. free of jealous tendencies
Placid, shy, bashful vs. composed
Suspicious vs. trustful
Dour vs. cheerful
Rigid vs. adaptable
Hard and unconcerned vs. concerned about other people

....Objectively the individual reports that he is annoyed by people claiming to be superior to others and that he considers himself as a person who is scrupulously correct in his behavior. He also shows some intensive interests in internal, mental life.

26. Factor M. Bohemianism vs. Practical Concernedness, M

A score of 52 is possible for this factor. This factor loads:

Unconventional, eccentric vs. conventional
Sensitively imaginative vs. practical, logical
Undependable vs. conscientious
Placid exterior vs. easily concerned and expressive
Occasional hysterical emotion vs. given to keeping head in emergencies

....There are some evidences in the questionnaire of the conversion hysteria type of instability at the "Bohemian" pole of this factor. For example, the person walks and talks in his sleep, makes demands on others with placidity, gets emotional but does not worry, is oblivious to convention but can feel childishly dependent and insecure episodically.

27. Factor N. Sophistication vs. Rough Simplicity, N

A score of 40 is possible for this factor. This factor loads:

Polished vs. clumsy, awkward
Cool, aloof vs. attentive to people
Fastidious vs. easily pleased

....In terms of the questionnaires, it has been discovered two or three times under varying titles such as intellectual leadership, hard headed rationalism, and titles indicating a sophisticated, intellectual, unsentimental approach to things.

28. Factor O. Worrying Suspiciousness vs. Calm Trustfulness, O

A score of 52 is possible for this factor. This factor is not very well defined in ratings, but loads:

Worrying, anxious vs. placid, tough
Suspicious, brooding vs. trustful, free from suspiciousness

In various questionnaires it has been very well defined and variously labelled as "Depressive Tendency," "Moodiness," "Emotional Sensitivity," "Self Depreciation," and "General Neuroticism." It is, however, quite distinct from the general neuroticism factor labelled C above, and it may be that it is some relatively temporary condition of depression and worrying which is not a stable feature of the personality. At any rate, it is somewhat decidedly more obvious in the mental interior of the subject than it is to the observer, causing him to feel remorseful, downhearted, subject to phobias and neurasthenic symptoms, worrying, avoiding people and perturbed by the inconstancy of his own moods.

29. Factor Q_1. Radicalism vs. Conservatism, Q_1

A score of 40 is possible for this factor. This factor has not appeared in personality ratings, and appears most clearly in attitudes. Nevertheless, it does have some role in the total personality, and the more radical show evidence of being more introspective, more interested in fundamental issues, and more interested in intellectual matters generally. There is also some evidence from more objective test situations that Q_1+ persons are more well informed, less inclined to moralize, and more inclined to experiment in life generally. The views of the more radical or more conservative on a set of actual attitudes are already clearly indicated in the questionnaire itself. This factor and the remaining three factors are not given alphabetical labels because they have not yet been demonstrated to appear in the series of factors obtainable in ratings.

30. Factor Q_2. Independent Self-Sufficiency vs. Lack of Resolution, Q_2

A score of 40 is possible for this factor. As will be evident from the questionnaire responses themselves, this factor in its positive loadings indicates an individual who is resolute and accustomed to going his own way, but who is not necessarily dominant in his relation to other people. The Q_2- person prefers to work and make decisions in company with other people, likes social approval and admiration, is conventional and fashionable. The predictive value will be evident from the occupational associations.

31. Factor Q_3. Will Control and Character Stability, Q_3

A score of 40 is possible for this factor. This factor has some relation to the C and G factors described above, yet it is not very evident in ratings and is still listed and defined primarily as a factor in the questionnaire responses for mental inferiors. Individuals high in this factor show, according to the questionnaire responses, strong control of emotions and behavior. They are inclined to be considerate, careful, and conscientious, but also obstinate. Although no behavior ratings have picked out the factor, it has been observed that when people who are high in this factor have been put together in groups, the general efficiency and objectiveness of the group is decidedly higher than in groups that are only average in the factor. There are indications that persons high in this factor are inclined to mathematical interests.

32. Factor Q_4. Nervous Tension, Q_4

A score of 52 is possible for this factor. This factor has been found in earlier questionnaire studies where it has been called nervousness, nervous anxiety and instability, and sleep difficulties with somaesthenia. The general picture is that of a person who is tense, fatigued but is unable to remain inactive. The dimension resembles a common description of what distinguished the hypertensive person. The ratings show some correlation between this factor and strength of interest in the opposite sex, and it is possible that it is to be explained as a sex suppression factor. Until the research now in progress is completed, however, the predictions from this factor must rest on the known character of the responses, and on whatever associations with occupational norms are shown below.

Personality of the teachers used in this study will be assessed by means of the Cattell Sixteen Personality Factor Questionnaire. Cattell makes this statement regarding this questionnaire:[15] "The questionnaire thus aims to leave out no important aspect of the total personality, for the above factors are based on even sampling from the personality sphere and include abilities (intelligence), temperamental factors, and dynamic (character integration) source traits." Cattell[16] makes the following distinctions between the three modalities of dynamic traits (or interests), abilities, and temperamental traits: "Dynamic traits are characterized by behavior arising from a stimulus situation or incentive and directed to some goal, at which the action ceases. Abilities, by contrast, are shown by how well the person makes his way to the accepted goals. Dynamic traits are thus traits in which performance varies as the incentive varies; whereas abilities can be recognized as those in which performance varies in response to changes in complexity. Temperament traits are definable by exclusion as those traits which are unaffected by incentive or complexity. These are traits like high-strungness, speed, energy, and emotional reactivity, which common observation suggests are largely constitutional."

The first twelve primary source traits of personality described earlier in this section have been determined by previous research, primarily factor analysis, to be shown in behavior ratings, questionnaire or "self-rating" items, and in objective tests. The other four primary personality factors have thus far been found only in questionnaire responses.

33. Practice Teaching Grade in Education 75, PT75

This grade is available on a special card filed in the education office at Wisconsin High School. Since such letter grades usually carry plus and minus signs the following transformation scheme was used: A = 1; A- = 2; B+ = 3; B = 4; C- = 8. A low score indicated a high grade.

---

[15] R. B. Cattell and others. Handbook for the Sixteen Personality Factor Questionnaire (Champaign, Illinois: Institute for Personality and Ability Testing), p. 1.

[16] Raymond B. Cattell. Personality (New York: McGraw-Hill Book Co., 1950), p. 35.

34. Practice Teaching Grade in Education Methods Course, PTEM

This grade is also filed at Wisconsin High School. The letter grades were converted into numerical scores using the same procedure as that outlined in 33, above. A low score indicates a high grade.

35. Grade in Education 73, G73, and
36. Grade in Education 74, G74

These grades were obtained from the undergraduate transcript of credits of these teachers. The letter grades were converted to numerical scores by using the following system: A = 1; B = 2; C = 3; and D = 4. A low score indicates a high grade.

37. Lecture Section Grade, Education 75, LG75

The 3-credit lecture section grade is combined with the 2-credit practice teaching grade to form the 5-credit Education 75 grade found on the University records. The 3-credit grade was determined by consultation with the individual instructors to obtain the best estimate which when weighted three times with twice weighted practice teaching would yield the 5-credit grade. A conversion system similar to that used for variables 35 and 36 was used to convert to numerical scores. A low score indicates a high grade.

38. Percentile Rank in High School Class, PRHS

The rank of each student in his high school graduating class was divided by the number of students in that class. This quotient was subtracted from 100 to yield the percentile rank in high school class used in this study. A high score indicates a high relative rank in high school graduating class.

39. American Council on Education Psychological Examination for College Freshmen Percentile Rank, ACEPR

The total score percentile rank was the measure of intelligence used in this study. Those teachers who entered the University of Wisconsin in the fall of 1946 were administered 1945 editions of this examination during Orientation Week, 1946. The majority of scores used in this study are from the 1945 edition of this test. National norms were used to determine the percentile ranks. The same norms were used for the several groups. A high score indicates a high level of intelligence.

40. Freshman-Sophomore Grade Point Average, Fr.-So. GPA

The following system was employed in the computation of this grade point average. A student may register as a junior in the School of Education at the University of Wisconsin if he has earned a minimum of 58 credits and an equivalent number of grade points. The majority of the students will register for their fifth semester as juniors. In such cases the number of grade points was divided by the number of credit hours to obtain the grade point average. Several cases did not have 58 credits after four semesters at the University. In these instances the grade point average for the fifth semester was computed. A sufficient number of credits and grade points of fifth semester were added to yield the minimum of 58 Freshman-Sophomore credits. The fifteen elective war service credits were not included for grade point ratio computations. They were, however, added to the credits a student had earned to change the student's University classification. If, for example, a student had earned 28 credits as a freshman and during the third semester he was granted these 15 war elective credits in addition to 15 credits of regular work, he would be permitted to register for his fourth semester of University study as a junior. But the freshman-sophomore grade point average would be computed only for the 43 credits he had earned at the University in regular class work. Transfer credits from other collegiate institutions were not included in the computation of GPA's. A high score indicates a high grade point average.

41. Junior-Senior Grade Point Average, Jr.-Sr. GPA

The credits and grade points, exclusive of war elective service credits, not included in the computation of the Freshman-Sophomore Grade Point Average were used to compute the Junior-Senior Grade Point Average. A high score indicates a high grade point average.

42. Composite: 2 Credit Major Educational Methods Lecture Section Grade and the 2 Credit Minor Methods Grade, EM4

The two credit major methods lecture section grade was obtained from the individual instructors in a manner similar to that outlined for the determination of the lecture grade in Education 75. The 2 credit minor methods grade was obtained from the University transcript. Each of these grades were converted to numerical scores with A = 1; B = 2; C = 3; D = 4. The sum of these two numerical scores was the score used for this composite measure. A low score indicates a high composite grade.

Procedure to be Followed in Treatment of Data

Success in teaching is defined in terms of the following ratings:

1. Principal's acceptability rating
2. Principal's M-blank rating (1st year)
3, 4. Two visiting supervisor's M-blank rat-

ings.

5. Principal's M-blank rating (2nd year).
6. A rating obtained from an outside agency on a special form.
7. Teacher's self evaluation.
8. Peer evaluation.
9. Pupil evaluation.

Four types of data were collected for each subject. They were: a) in-service estimates of success; b) measures of temperament; c) measures of personality; and d) pre-service measures of achievement. This investigation is primarily concerned with the relationships between a) and b), a) and c), and a) and d).

This study will follow the following procedure:

1. The coefficients of correlation for each pair of in-service estimates of success variables will be computed.
2. The resultant nine by nine matrix of correlations will be factored by the multiple group method to determine factors underlying these estimates of teaching success. These hypothetical factors should provide us with a simpler explanation of the interrelationships of the teaching success estimates.
3. The correlation coefficients of each temperament measure, each personality measure, and each achievement measure with each of the nine estimates of teaching success will be computed. These correlations will form the entries of three rectangular matrixes of orders seven by nine, sixteen by nine, and ten by nine respectively.
4. The grouping of the variables and a procedure similar to that utilized in (2) will be applied to these three rectangular matrixes to determine if such an analysis will account for the correlations in these matrixes.
5. Factor scores for the factors found in (2) will be calculated for each individual. These scores are estimates of the factors found by the analysis in (2). These factor scores will be used as criteria of teaching success.
6. Pearson product-moment coefficients of correlation will be calculated for each temperament, personality, and achievement variable with these obtained criteria.
7. Multiple regression techniques will be employed to develop equations for the prediction of the criteria.

SECTION III

ANALYSIS OF THE DATA

IN THIS section the statistical treatment of the data will be described. The readers will be referred to tables and to selected references for more detailed discussions of the statistical methods employed in carrying out this study.

The extreme scores, ranges, means, standard deviations, and the number of teachers for whom the forty-two measures were obtained are presented in Table II.

The Pearson product moment coefficient of correlation will be used in this study to indicate linear relationships existing between various sets of scores.

Factor analysis will be the major statistical device utilized in summarizing the correlational data.

This section will be divided into six parts. Part one will deal with the factor analysis of the nine by nine matrix of intercorrelations of the nine estimates of teaching success. This nine by nine matrix has been designated the R matrix.

In part two the analysis of the R_X matrix (seven by nine rectangular matrix) will be presented. This matrix has as entries the correlations of the seven temperament variables, numbered in this study 10 through 16, with the nine estimates of teaching success.

In part three the factor analysis of the rectangular matrix, of order sixteen by nine, designated the R_Y matrix, will be discussed. The entries in this matrix are the correlations of the sixteen personality source trait measures (variables 17 through 32) with the nine estimates of teaching success.

Part four will be concerned with a discussion of the factor analysis of the R_Z matrix. This matrix is of order ten by nine and has as entries the correlations of ten preservice measures of achievement and intelligence (variables numbered 33 through 42) with the nine in-service estimates of the factors found in the analysis reported in part one. These estimates will be in the form of factor scores and will be used as "composite" criteria in the prediction phase of this report.

Part six will discuss the results of the multiple correlation techniques employed to predict the factor scores discussed in the preceding part.

TABLE II

EXTREMES, RANGES, MEANS, STANDARD DEVIATIONS, AND NUMBER OF CASES OF THE FORTY-TWO MEASURES

Variable	Code	High* Score	Low Score	Range	Mean	S. D.	No.
1	PA	5	1	4	2.05	0.80	64
2	PM1	5	1	4	2.16	0.87	61
3	S1M	5	1	4	2.53	0.84	62
4	S2M	5	1	4	2.53	0.95	58
5	PM2	4	1	3	2.29	0.63	63
6	SpR	12	5	7	7.12	1.86	60
7	SE	3	1	2	2.18	0.66	62
8	PE	8	2	6	4.02	1.45	56
9	PuE	18	5	13	10.51	3.02	53
10	A-T	18	5	13	12.20	3.34	64
11	V-T	18	2	16	10.83	4.38	64
12	I-T	19	5	14	11.75	3.26	64
13	D-T	20	4	16	12.52	4.45	64
14	E-T	19	6	13	12.13	3.98	64
15	S-T	19	5	14	13.58	3.14	64
16	R-T	19	1	18	8.22	3.08	64
17	A	34	8	26	22.56	5.46	64
18	B	23	14	9	19.69	2.08	64
19	C	48	26	22	36.84	5.00	64
20	E	36	6	30	21.86	7.01	64
21	F	50	8	42	29.78	8.38	64
22	G	36	16	20	25.89	4.66	64
23	H	50	16	34	37.95	8.51	64
24	I	30	4	26	17.16	4.85	64
25	L	28	2	26	15.81	5.81	64
26	M	30	8	22	19.16	5.40	64
27	N	34	12	22	22.95	3.90	64
28	O	30	0	30	10.47	8.30	64
29	Q_1	30	8	22	19.31	5.58	64
30	Q_2	31	8	23	17.38	5.98	64
31	Q_3	38	10	28	26.55	5.64	64
32	Q_4	34	2	32	13.33	7.75	64
33	PT75	7	1	6	2.89	1.74	64
34	PTEM	8	1	7	2.61	1.78	64
35	G73	3	1	2	1.84	0.57	63
36	G74	3	1	2	1.90	0.53	63
37	LG75	4	1	3	2.17	0.78	64
38	PRHS	99	15	84	75.70	23.81	60
39	ACEPR	100	8	92	60.30	25.31	56
40	Fr-SoGPA	2.80	0.97	1.83	1.78	0.51	58
41	Jr-SrGPA	2.88	1.02	1.86	1.99	0.41	64
42	EM4	12	5	7	3.42	1.00	64

*High refers to numerical value only.

The Use of Factor Analysis

Before proceeding to the more detailed discussion of the findings of this study, the writer feels that several brief statements concerning factor analysis should be made.

Holzinger and Harman[17] make this statement about factor analysis: "Factor analysis is a branch of statistical theory concerned with the resolution of a set of descriptive variables in terms of a small number of categories or factors. This resolution is accomplished by the analysis of the intercorrelations of the variables. A satisfactory solution will yield factors which convey all the essential information of the original set of variables. The chief aim is thus to attain scientific parsimony or economy of description."

This is the first basic problem with which factor analysis deals. The second basic problem deals with the problem of describing the factors in terms of the original variables.

There are many different methods which can be used to transform a correlation matrix into a factor matrix. The multiple-group method of factoring the correlation matrix was chosen for use in this study. This method was first presented by Holzinger[18] and six months later by Thurstone.[19] This procedure is applicable to any correlation matrix which can be divided into sections that are linearly independent.

Factor analysis in this study was used both to determine the factors common to the nine estimates of teaching success and to determine whether these common factors underlie certain temperament, personality, and pre-service measures of achievement and intelligence. The results of these analyses will be reported in parts one through four. The second basic problem of factor analysis involving factor estimates will be discussed in part five.

The multiple-group method outlined by Harris and Schmid[20] was closely followed in this study.

Part One: Factor Analysis of the Nine Inservice estimates of Teaching Success

The writer doubted if these nine estimates were measures of the same factor. It was found that a general factor would not account for the intercorrelations of the R matrix supporting the original contention. It seemed desirable to obtain a simpler description of these measures. That an arithmetical average of the measures in the form of a composite might not be a satisfactory means of summarizing this data has been pointed out by Holzinger.[21]

Factor analysis of the nine in-service estimates of teaching success was designed to follow these steps:

1. Calculation of the matrix of intercorrelations of the nine estimates of teaching success.
2. Transformation of this correlation matrix into an arbitrary oblique factor matrix using the multiple group method of factoring.
3. Rotation of this initial oblique solution to a meaningful primary pattern following the method of direct rotation to primary factor structure outlined by Harris.[22]

It became evident after the completion of the first two steps that the third step would not be necessary.

Pearson product-moment coefficients of correlation were computed for each pair of the first nine variables. The intercorrelations and the final communality estimates of these variables are presented in Table III. This matrix has been labeled the R-matrix for subsequent references.

After several attempts at grouping the variables and using Harris'[23] geometrical criterion of the dimensionality of the common factor space, R was finally sectioned into three groups as follows:

I:	1, 2, 3, 4
II:	5, 6, 7
III:	8, 9

---

[17] Karl J. Holzinger and Harry H. Harmon. Factor Analysis: A Synthesis of Factorial Methods (Chicago: University of Chicago Press, 1948), p. 3.

[18] Karl J. Holzinger. "A Simple Method of Factor Analysis," Psychometrika, IX (December 1944), pp. 257-261.

[19] L. L. Thurstone. "A Multiple Group Method of Factoring the Correlation Matrix," Psychometrika, X (June 1945), pp. 73-78.

[20] Chester Harris and John Schmid. "Further Application of the Principles of Direct Rotation in Factor Analysis," Journal of Experimental Education, XVIII (March 1950), pp. 175-193.

[21] Karl J. Holzinger. "Factoring Test Scores and Implications for the Method of Averages," Psychometrika, IX (September 1944), pp. 155-167.

[22] Chester W. Harris. "Direct Rotation to Primary Structure," Journal of Educational Psychology, XXXIX (December 1948), pp. 449-468.

[23] Chester W. Harris. "Projections of Three Types of Factor Pattern," Journal of Experimental Education, XVII (March 1949), pp. 335-345.

TABLE III

INTERCORRELATIONS, COMMUNALITY ESTIMATES AND RESIDUAL MATRIX FOR NINE INSERVICE ESTIMATES OF TEACHING SUCCESS

R MATRIX

	1	2	3	4	5	6	7	8	9	h²
1		07*	-06	-03	-01	00	02	-01	01	75
2	77 (61)**		-03	-05	02	09	-10	-02	02	70
3	57 (62)	57 (59)		06	01	-03	02	04	-04	60
4	65 (58)	61 (56)	68 (58)		-02	-05	07	-01	01	70
5	43 (63)	41 (60)	29 (61)	29 (57)		-03	00	-08	08	65
6	43 (60)	49 (57)	24 (58)	25 (54)	54 (59)		00	08	-08	60
7	17 (62)	03 (59)	05 (60)	11 (56)	38 (61)	37 (58)		00	00	30
8	18 (56)	19 (55)	09 (54)	06 (51)	28 (55)	50 (55)	33 (54)		-01	50
9	08 (53)	15 (53)	-07 (51)	01 (49)	24 (53)	17 (49)	22 (51)	41 (48)		45

* Decimal points omitted.
Residuals have been placed in upper triangle, underlined.

**Number inside parenthesis indicates the size of the sub-group used to calculate the correlation.

The correlations of each variable were summed over the three groups to produce G. In this case G is a nine by three matrix. The above operation is equivalent to post-multiplying R by a matrix E which has zero and unity elements arranged as follows:

TABLE IV

E MATRIX

	I	II	III
1	1	0	0
2	1	0	0
3	1	0	0
4	1	0	0
5	0	1	0
6	0	1	0
7	0	1	0
8	0	0	1
9	0	0	1

A three by three matrix, H, was obtained by summing the columns of G over each group. This matrix represents the correlations of each group with every other group. In matrix notation the above operations can be summarized as follows:

$$RE = G$$
$$E'RE = H$$

Next a diagonal matrix W was constructed. This matrix has as elements the reciprocal square roots of the diagonal elements of H.

The G matrix, the H matrix, and the W matrix are shown in Tables V, VI, and VII, respectively. Pre- and post-multiplication of H by W yields Φ, the matrix of intercorrelations of the group axes. This matrix is shown in Table VIII.

The structure matrix S, is shown in Table IX. This matrix was obtained by post-multiplying G by W, and represents the correlations of each variable with the oblique group factors. Post-multiplication of S by Φ^{-1} yields the matrix P. P is the matrix of coordinates of the variables with respect to these factors. The matrix P is shown in Table X. The values in P will be called factor loadings in the remainder of this study.

The matrix multiplication $SP = R+$ yields the nine by nine matrix of reproduced correlations. One should note that the matrix multiplication $GH^{-1}G'$ will also yield this matrix of reproduced correlations. The matrix subtraction $R - R+$ was performed to yield the residual correlations shown in the upper triangle of Table III. An examination of these residuals reveals that the calculated oblique solution reproduced the original correlations very well.

It was not considered necessary to rotate this oblique solution to another oblique solution since the pattern matrix, P, indicated that each of the variables are of complexity one.

Inferences concerning the nature of the factors were made in terms of the number of significant loadings in each column of the pattern matrix and the complexities of the variables. A variable is said to be of complexity one if it has only one significant loading in the pattern matrix. In this study a minimal loading of .30 was required for a factor to represent a contribution to the variance of a test. Thurstone's[24] criteria for uniqueness of simple structure were used as a basis of interpretation of the pattern values obtained in this study. The pattern values obtained, without resorting to further rotation and change of reference axes, exhibit the properties of simple structure outlined by Thurstone. This led the investigator to conclude that this solution does not differ greatly from a primary axis solution since it is the primary pattern in such a solution which will exhibit Thurstone's properties of simple structure.

The tentative names and the factor loadings on each of these oblique factors were as follows:

I. Beginning Teacher Rating Scale Factor

1. PA, Principal's Acceptability Rating	.783
2. PM1, Principal's M-Blank rating, 1st year	.811
3. S1M, Supervisor I M-blank rating	.782
4. S2M, Supervisor II M-blank rating	.858

These four variables have significant factor loadings on this factor. Each of the variables is of complexity one. The highest loadings are for variables 2 and 4. Variables 1 and 3 have practically identical loadings on this factor.

II. Second-year Rating Scale Factor

5. PM2, Principal's M-blank rating, 2nd year	.858
6. SpR, Outside Agency Rating	.641
7. SE, Teacher's Self Evaluation	.536

Variables 5, 6, and 7 have significant loadings on this factor. Variable 5 has the highest loading. Each of these variables is of complexity one.

III. Peer-Pupil Response Factor

8. PE, Peer evaluation	.561
9. PuE, Pupil Evaluation	.769

Variables 8 and 9, both of complexity one, have significant loadings on this factor. The

---

[24] L. L. Thurstone. *Multiple-Factor Analysis* (Chicago: University of Chicago Press, 1947), p. 335.

TABLE V

CORRELATIONS OF EACH VARIABLE WITH EACH GROUP G MATRIX

	I	II	III
1.	2.74	1.03	0.26
2.	2.65	0.93	0.34
3.	2.42	0.58	0.02
4.	2.64	0.65	0.07
5.	1.42	1.57	0.52
6.	1.41	1.51	0.67
7.	0.36	1.05	0.55
8.	0.52	1.11	0.91
9.	0.17	0.63	0.86

TABLE VI

H MATRIX

	I	II	III
I	10.45	3.19	0.69
II	3.19	4.13	1.74
III	0.69	1.74	1.77

TABLE VII

W MATRIX

	I	II	III
I.	.3093	0000	0000
II.	0000	.4921	0000
III.	0000	0000	.7517

TABLE VIII

ϕ MATRIX

	I	II	III
I.	1.000	.486	.160
II.	.486	1.000	.644
III.	.160	.644	1.000

TABLE IX

CORRELATIONS OF EACH VARIABLE WITH EACH OBLIQUE FACTOR
S MATRIX (STRUCTURE MATRIX)

	I	II	III
1.	848*	507	195
2.	820	458	256
3.	749	285	015
4.	817	320	053
5.	439	773	391
6.	436	743	504
7.	111	517	413
8.	161	546	684
9.	053	310	646

*Decimal points omitted

TABLE X

COORDINATES OF EACH VARIABLE WITH EACH FACTOR
P MATRIX (PATTERN MATRIX)

	I	II	III
1.	783*	140	-021
2.	811	-029	144
3.	782	-041	-084
4.	858	-072	-038
5.	050	858	-169
6.	113	641	073
7.	-165	536	094
8.	-025	197	561
9.	025	-197	769

*Decimal points omitted

highest loading (.769) is for variable 9.

Summary of the Factor Analysis of Teaching Success Estimates

It was found that a general factor would not account for the matrix of intercorrelations of the nine estimates of teaching success.

Three common factors identified as

I. Beginning Teacher Rating Scale Factor,
II. Second-Year Rating Scale Factor,
III. Peer-Pupil Response Factor

were found in the analysis described above. The factors are oblique with sizeable correlations between Beginning Teacher Rating Scale Factor and Second-Year Rating Scale Factor (.49); Second-Year Rating Scale Factor and Peer-Pupil Response Factor (.64). The correlation between Beginning Teacher Rating Scale Factor and Peer-Pupil Response Factor was low (.16). All of the correlations are positive.

From the above analysis one must conclude that an arithmetic average of these nine estimates of teaching success would not yield a satisfactory composite estimate.

In the fifth part of this section, estimates of these factors in the form of factor scores will be obtained and used as "composite" criteria of teaching success.

Part Two: The Factor Analysis of the R_X Matrix

The correlations of each of the seven temperament traits with each of the nine estimates of teaching success were calculated. These correlations are the entries in the seven by nine matrix labeled the R_X Matrix and recorded in Table XI. Each of these temperament variables had been found, in previous research studies, to be of complexity one. An examination of the R_X Matrix reveals that 18 of the 63 correlations are negative. The 63 correlations ranged from -.26 (N = 58, between Thurstone Active, 10, and Supervisor II M-blank rating, 4) to .31 (N = 58, between Sociable, 15, and Supervisor II M-blank rating, 4). For a random sample of 64, the standard error of a zero correlation is .125. Thus, a coefficient of .25 or greater is necessary in order for the correlation to be significantly different from zero at the five percent level. For 58 cases a coefficient of .26 or larger is necessary for statistical significance at the same level. The 64 subjects used in this study do not represent a random sample. However, it was decided to use the above levels of significance in analyzing the data. Only the two correlations reported above and the correlation of .26 (N = 64, between Dominant, 14, and Acceptability, 1) are significantly different from zero by the above mentioned standards. In addition to these three correlations, seven other correlations are of magnitude .20 or larger. Temperament variables numbered 11, 13, 14, and 15 had correlations of .20 or larger with teaching success estimate 6. In general these correlations tend to be quite small. This leads one to conclude that there is very little linear relationship between these seven temperament traits (as assessed by the Thurstone Temperament Schedule) and the nine estimates of teaching success.

Factor analysis techniques were employed in analyzing the correlations in the R_X matrix (and also the R_Y and R_Z matrices) to determine if the three oblique common factors described in part one underlie temperament, personality, and pre-service measures of achievement and intelligence. If these factors do underlie such measures then an analysis similar to that employed in the analysis of the original nine by nine matrix should account for the correlations in these three rectangular matrixes. If the residual correlations are small, then an examination of the appropriate pattern matrixes will reveal the temperament, personality, and pre-service achievement variables which have significant loadings on the three oblique teaching success factors. Such a procedure seemed to be a very straight forward and sensible one since in this study we are concerned only with the relationships between these temperament, personality and achievement variables and estimates of teaching success. Thus, for example, we were not primarily concerned with the intercorrelations of the temperament variables nor with the correlations of the temperament variables with the personality variables.

The R_X Matrix was postmultiplied by the E matrix (Table III) to yield the G_X matrix (Table XII). The matrix multiplication $G_X H^{-1} G'$ was performed resulting in a reproduced correlation matrix of order seven by nine labeled R_X^+. This latter matrix was subtracted from the R_X matrix to yield a matrix of residual correlations ($R_X - R_X^+$). This matrix may be found in Table XIII. One notes that the residuals with few exceptions are very small. It thus appears that the three teaching success factors do underlie these seven temperament traits. The correlations of the temperament variables with these oblique factors may be found in Table XIV. They were obtained by the matrix multiplication $G_X W = S_X$ (Structure Matrix). The P_X matrix (Pattern Matrix) showing the coordinates of the seven temperament variables with respect to the oblique factors is recorded in Table XV. This matrix was obtained by post-multiplying S_X by Φ^{-1}.

It is interesting to note that the correlations of the seven temperament variables with the three oblique factors are small. Variables 11

TABLE XI

CORRELATIONS OF SEVEN TEMPERAMENT TRAITS WITH NINE ESTIMATES OF TEACHING SUCCESS

R_X MATRIX

	1 (64)*	2 (61)	3 (62)	4 (58)	5 (63)	6 (60)	7 (62)	8 (56)	9 (53)
10	-17**	-04	-14	-26	-14	15	-14	11	21
11	12	19	-08	-03	05	24	-09	16	05
12	18	01	-02	07	15	16	-07	06	-12
13	26	16	06	17	09	22	12	08	-12
14	16	16	08	09	06	20	09	04	00
15	23	18	09	31	12	20	12	-01	-19
16	08	-01	20	01	-01	03	16	00	-10

* Number inside parenthesis indicates the size of the subgroup used to calculate the correlations in that column.

**Decimal points omitted.

TABLE XII

CORRELATIONS OF EACH TEMPERAMENT VARIABLE WITH EACH GROUP

G_X MATRIX

	I	II	III
10	-61*	-13	32
11	20	20	21
12	24	24	-06
13	65	43	-04
14	49	35	04
15	81	44	-20
16	28	18	-10

*Decimal points omitted

TABLE XIII

RESIDUAL CORRELATIONS

$R_X - R_X^+$ MATRIX

	1	2	3	4	5	6	7	8	9
10	-01*	08	03	-09	-04	19	-16	-02	02
11	06	12	-11	-07	-01	16	-15	05	-05
12	11	-04	-08	02	04	08	-12	06	-06
13	07	01	-09	01	-11	06	04	06	-06
14	02	04	-03	-03	-09	07	02	-01	01
15	00	00	-11	10	-10	04	06	04	-04
16	00	-07	13	-06	-10	-03	13	03	-03

*Decimal points omitted

11.8.70
g 107

TABLE XIV

CORRELATIONS OF EACH TEMPERAMENT VARIABLE WITH EACH OBLIQUE FACTOR

S_X MATRIX (STRUCTURE MATRIX)

	I	II	III
10	-189*	-064	241
11	062	098	158
12	074	118	-045
13	201	212	-030
14	152	172	030
15	251	217	-150
16	087	089	-075

*Decimal points omitted

TABLE XV

CORRELATIONS OF EACH TEMPERAMENT VARIABLE WITH RESPECT TO EACH OBLIQUE FACTOR

PX MATRIX (PATTERN MATRIX)

	I	II	III
10	-117*	-298	451
11	054	-041	176
12	-020	264	-212
13	076	345	-264
14	066	218	-121
15	097	472	-469
16	013	226	-222

*Decimal points omitted

and 14 are positively correlated with each oblique factor axis.

None of the seven variables have significant factor loadings on Factor I. Variables 13 and 15 have significant loadings on Factor II while variables 10 and 15 have significant loadings on Factor III.

It should be noted that R_X^+ can also be obtained by following either one of the following matrix multiplications:

$$S_X P' = R_X^+ \quad \text{or} \quad P_X S' = R_X^+ .$$

The reproduced correlations of the R_y and the R_z matrixes (to be discussed in the next two parts) can also be computed by substituting in the above equations the structure and pattern values associated with a particular matrix.

Part Three: The Factor Analysis of the R_y Matrix

The correlations of each of the sixteen personality variables with each of the nine estimates of teaching success were calculated. These correlations are the entries in the sixteen by nine rectangular matrix labeled the R_y Matrix and recorded in Table XVI.

An examination of the R_y Matrix reveals that over half of the correlations are negative (75 of the 144 correlations). The sign of each of these correlations was changed so that they would agree with other correlations, as regards sign and meaning. (A low number for the inservice estimates indicates a high rating whereas a high number for the personality factor scores indicates the plus pole of the given trait.) The 144 correlations ranged from -.32 (N = 53, between Q_1, 29, and Pupil Evaluation, 9) to .38 (N = 38, between Q_3, 31, and Self-Evaluation, 7). In addition to these two correlations, ten correlations were significantly different from zero by the standards discussed in Section 2. Including these twelve correlations, there are a total of 34 correlations of .20 or larger in the matrix. All of the correlations of variable 26 (M) with the estimates of teaching success are negative. Seven of the nine correlations of variable 28 (0) with these estimates are negative. All nine correlations of variable 31 (Q_3) with these estimates are positive with seven of them being .20 or larger. Six of the nine correlations of variable 22 (G) with the nine teaching success estimates are .20 or larger. All nine correlations are positive. The 45 correlations of variables 17 (A), 19 (C), 21 (F), 25 (L), and 30 (Q_2) with the teaching success estimates are all less than .20. All of the correlations tend to be quite small. Therefore, there seems to be little relationship between these particular estimates of teaching success and the sixteen source traits of personality (as assessed by the 16 P. F. Test).

The R_y Matrix was post-multiplied by the E Matrix to yield the G_y Matrix (Table XVII). The matrix multiplication $G_yH^{-1}G'$ yielded a reproduced correlation matrix of order sixteen by nine labeled R_y^+. The matrix subtraction $R_y - R_y^+$ yielded the matrix of residual correlations ($R_y - R_y^+$). (See Table XVIII.) These residuals tend to be quite small. It thus appears that the factoring method employed accounts for these correlations. Therefore one might conclude that the three oblique factors also underlie the sixteen personality source trait variables. The correlations of the personality variables with these oblique factors may be found in Table XIX. This sixteen by three structure matrix was obtained by the matrix multiplication $G_yW = S_y$. The pattern matrix P_y was obtained by the matrix multiplication $S_y\Phi^{-1}$. This matrix will be found in Table XX and lists the coordinates of the sixteen personality variables with respect to the three oblique factors. These coordinates as pointed out earlier in this section are also called the factor loadings.

The structure matrix S_y reveals that the correlations of the sixteen personality variables with the three oblique factors are low. Of these sixteen variables only variables 22 and 31 are positively correlated with each oblique axis.

An examination of Table XX indicates that variables 18 and 29 have significant loadings on Factor I. On Factor II will be found significant loadings for ten of the sixteen personality measures. The variables having significant loadings are 17, 18, 20, 23, 24, 27, 28, 29, 31, and 32. Variables 20, 22, 23, 24, 26, and 29 have significant loadings on Factor III.

Part Four: The Factor Analysis of the R_z Matrix

The correlations of each of the ten achievement variables with each of the nine estimates of teaching success were calculated and recorded in the ten by nine R_z Matrix (Table XXI).

Thirty of these ninety correlations are negative. It is interesting to note that all nine correlations of variable 33 (PT75) with the nine estimates are positive with four of them being significant at the 5% level of significance. Likewise all of the correlations of variable 34 (Practice Teaching Grade in Educational Methods Course) with these nine estimates of success, although small, are positive. These correlations range from -.29 (N = 52, between variable 35, Grade in Education 73, and pupil evaluation, variable 9) to .28 (N = 62, between variable 33, PT75, and variable 2, PM1). Only ten of the correlations are .20 or larger. Eight of the nine correlations of variable 36 with these estimates of teaching success are negative.

TABLE XVI

CORRELATIONS OF SIXTEEN PERSONALITY FACTORS WITH NINE ESTIMATES OF TEACHING SUCCESS

R_Y MATRIX

	1 (64)*	2 (61)	3 (62)	4 (58)	5 (63)	6 (60)	7 (62)	8 (56)	9 (53)
17	-13**	00	01	14	15	06	08	-03	-04
18	-12	-20	-21	-20	09	03	28	07	22
19	01	-05	-08	06	-03	-05	01	08	-11
20	-05	05	-04	-04	-06	20	05	-13	-02
21	00	00	03	16	00	07	-10	-09	-12
22	10	06	26	23	22	06	22	21	27
23	11	14	-11	09	16	36	01	07	-14
24	09	-18	07	-06	-11	-06	06	-17	-23
25	-06	-10	-01	-15	03	-03	-03	-05	05
26	-09	-29	-17	-16	-23	-12	-04	-23	-28
27	-21	04	-11	-03	12	23	18	09	25
28	-16	-10	00	-15	-12	-27	-27	-17	08
29	-17	-14	-07	-21	-08	00	17	-11	-32
30	-01	01	10	-19	-11	00	-02	02	-12
31	28	20	09	20	28	21	38	20	00
32	19	-18	-03	-25	-13	-13	-22	-10	12

---

* Number inside parenthesis indicates the size of the subgroup used to calculate the correlations in that column.

**Decimal points omitted.

TABLE XVII

CORRELATIONS OF EACH PERSONALITY VARIABLE WITH EACH GROUP

G_Y MATRIX

	I	II	III
17	02*	29	-07
18	-73	40	29
19	-06	-07	-03
20	-08	19	-15
21	19	-03	-21
22	65	50	48
23	23	54	-07
24	-08	-11	-40
25	-32	-03	00
26	-71	-39	-51
27	-31	53	34
28	-41	-66	-09
29	-59	09	-43
30	-09	-13	-10
31	77	87	20
32	-65	-48	02

*Decimal points omitted

TABLE XVIII

RESIDUAL CORRELATIONS

$R_Y - \overset{+}{R}_Y$ MATRIX

	1	2	3	4	5	6	7	8	9
17	-16*	01	01	14	02	-03	01	-03	03
18	03	-04	-01	02	-03	-09	12	-10	10
19	03	-03	-07	07	00	-02	03	10	-10
20	-05	09	-03	-02	-16	15	01	-09	09
21	-05	-03	-03	10	-02	08	-07	00	00
22	-08	-15	14	09	06	-15	09	-04	04
23	02	10	-16	05	-08	20	-12	05	-05
24	11	-12	06	-05	-11	00	11	01	-01
25	02	-02	07	-07	05	-01	-04	-05	05
26	10	-06	-04	00	-12	06	06	03	-02
27	-17	09	00	08	-06	06	00	-11	11
28	-02	00	08	-07	15	-04	-11	-07	07
29	-04	06	05	-06	-16	02	15	06	-06
30	02	04	11	-17	-07	05	02	07	-07
31	04	00	-07	03	-07	-10	18	04	-04
32	00	-03	12	-10	08	04	-13	-07	07

*Decimal points omitted

TABLE XIX

CORRELATIONS OF EACH PERSONALITY VARIABLE WITH EACH OBLIQUE FACTOR

S_Y MATRIX (STRUCTURE MATRIX)

	I	II	III
17	006*	143	-053
18	-226	197	218
19	-019	-034	-023
20	-025	093	-113
21	059	-015	-158
22	201	246	361
23	071	266	-053
24	-025	-054	-301
25	-099	-015	000
26	-220	-192	-383
27	-096	261	256
28	-127	-325	-068
29	-182	044	-323
30	-028	-064	-075
31	238	428	150
32	-201	-236	015

*Decimal points omitted

TABLE XX

COORDINATES OF EACH PERSONALITY VARIABLE WITH RESPECT TO EACH OBLIQUE FACTOR

P_Y MATRIX (PATTERN MATRIX)

	I	II	III
17	-140*	393	-284
18	-411	365	049
19	-004	-030	-003
20	-159	387	-336
21	038	123	-243
22	185	-098	394
23	-161	617	-424
24	-094	300	-479
25	-123	055	-015
26	-268	268	-512
27	-276	344	079
28	093	-541	265
29	-407	695	-705
30	-008	-022	-060
31	-003	568	-215
32	-059	-381	269

*Decimal points omitted

TABLE XXI

CORRELATIONS OF TEN PRE-SERVICE MEASURES OF ACHIEVEMENT WITH NINE ESTIMATES OF TEACHING SUCCESS

R_Z MATRIX

	1	2	3	4	5	6	7	8	9
33	08* (64)**	28 (61)	25 (62)	17 (58)	11 (63)	24 (60)	28 (62)	28 (56)	12 (53)
34	05 (64)	06 (61)	13 (62)	19 (58)	03 (63)	08 (60)	07 (62)	19 (56)	10 (53)
35	09 (63)	-16 (60)	06 (61)	00 (57)	00 (62)	-09 (59)	09 (61)	-07 (55)	-29 (52)
36	-02 (63)	-03 (60)	-03 (61)	-10 (57)	07 (62)	-07 (59)	-04 (61)	-10 (55)	-20 (53)
37	19 (64)	05 (61)	00 (62)	03 (58)	12 (63)	-04 (60)	12 (62)	06 (56)	-14 (53)
38	-07 (60)	-04 (57)	06 (58)	14 (54)	-17 (59)	-01 (58)	-06 (59)	-10 (53)	-17 (49)
39	08 (56)	-18 (53)	-10 (55)	06 (53)	-23 (55)	10 (52)	04 (55)	-09 (48)	-07 (46)
40	06 (58)	09 (56)	-08 (56)	-08 (53)	00 (57)	18 (54)	10 (56)	10 (51)	-19 (49)
41	10 (64)	11 (61)	07 (62)	-02 (58)	10 (63)	24 (60)	06 (62)	13 (56)	22 (53)
42	13 (64)	03 (61)	07 (62)	11 (58)	18 (63)	09 (60)	06 (62)	15 (56)	04 (53)

* Decimal points omitted.
**Number inside parenthesis indicates the size of the subgroup used to calculate the correlation.

The Educational Methods Composite Lecture Section Grade, Variable 42, correlates positively but low with these nine variables.

The R_Z Matrix was post-multiplied by the E Matrix to yield the G_Z Matrix (Table XXII). The ten by nine reproduced correlation matrix R_Z^+ was obtained by the matrix multiplication $G_ZH^{-1}G'$. This reproduced matrix was subtracted from R_Z to produce the $R_Z - R_Z^+$ matrix of residual correlations. This matrix may be found in Table XXIII. Since these residuals tend to be small one must conclude that the initial oblique solution applied to his matrix satisfactorily accounts for the correlations. The ten by nine structure matrix S_Z will be found in Table XXIV. S_Z was obtained from the matrix multiplication $G_ZW = S_Z$. This matrix shows the correlations of the ten preservice measures of achievement with the three oblique factors. The pattern matrix P_Z shown in Table XV was obtained by the matrix multiplication $S_Z\Phi^{-1} = P_Z$. These values give the coordinates of these ten variables with respect to the three oblique factor axes.

An examination of the S_Z matrix reveals that the correlations with the oblique axes are low. Variables 33, 34, 41, and 42 are positively correlated with each axis.

None of these ten variables have significant loadings on Factor I (refer to the P_Z matrix). Variables 35, 36, and 40 have small but significant loadings on Factor II and on Factor III significant loadings for variables 34, 35, 36, and 40 will be found.

The results of the factor analyses reported in the first four parts of this section could be summarized as follows:

1. Three oblique factors accounted for the matrix of intercorrelations of the nine in-service estimates of teaching success, and
2. These three factors appeared to underlie the temperament, personality, and achievement measures used in this study.

Part Five: The Factor Estimates

In the first four parts of this section we have been concerned with the problem of the linear resolution of the variables in terms of three hypothetical factors. In this part regression techniques will be employed to describe these three oblique factors in terms of the nine inservice estimates of teaching success. The weighting of each variable in such an estimation equation is determined partly by the correlation (structure value) of that variable with the factor.

The equations for estimating the three factors were determined by use of the shortened method presented by Holzinger and Harman.[25]

In standard form the equations are:

$$(1)\ \overline{F}_I = .312z_1 + .272z_2 + .192z_3 + .281z_4 + .042z_5 + .042z_6 - .018z_7 - .008z_8 - .008z_9$$

$$(2)\ \overline{F}_{II} = .127z_1 + .049z_2 - .015z_3 - .025z_4 + .407z_5 + .303z_6 + .159z_7 + .168z_8 + .047z_9$$

$$(3)\ \overline{F}_{III} = .000z_1 + .119z_2 - .081z_3 - .074z_4 + .052z_5 + .173z_6 + .108z_7 + .375z_8 + .396z_9$$

An estimated factor is not in standard form although such a variable has a mean of zero and a standard deviation equal to the coefficient of multiple correlation.[26]

The coefficients of multiple correlation for these factors were as follows:

$$\text{S.D. } \overline{I} = R_{\overline{I}} = .946$$

$$\text{S.D. } \overline{II} = R_{\overline{II}} = .896$$

$$\text{S.D. } \overline{III} = R_{\overline{III}} = .831$$

The estimation equations when expressed in terms of the observed values are:

$$(4)\ \overline{F}_I = .390x_1 + .313x_2 + .229x_3 + .296x_4 + .067x_5 + .023x_6 - .027x_7 - .006x_8 - .003x_9 - 3.01$$

$$(5)\ \overline{F}_{II} = .159x_1 + .056x_2 - .018x_3 - .026x_4 + .646x_5 + .163x_6 + .241x_7 + .116x_8 + .016x_9 - 4.14$$

$$(6)\ \overline{F}_{III} = .000x_1 + .137x_2 - .096x_3 - .078x_4 + .083x_5 + .093x_6 + .164x_7 + .259x_8 + .131x_9 - 3.48$$

Each of the above equations was evaluated for each teacher. The factor scores are shown in Table XXVI.

In order to eliminate negative values a transformation to an arbitrary positive scale was made.

The factor scores were standardized and this variable was equated to an arbitrary variable X with mean of 5 and standard deviation of 1. This transformation can be written:

$$(7)\qquad \overline{X}_{si} = \frac{\overline{F}_{si}}{R_{\overline{F}_s}} + 5$$

---

[25] Op. cit., pp. 278-83.

[26] Ibid., p. 273.

TABLE XXII

CORRELATIONS OF EACH ACHIEVEMENT VARIABLE WITH EACH GROUP

G_Z MATRIX

	I	II	III
33	78*	63	40
34	43	18	29
35	-01	00	-36
36	-18	-04	-30
37	27	20	-08
38	09	-24	-27
39	-14	-09	-16
40	-01	28	-09
41	26	40	35
42	34	33	19

*Decimal points omitted

TABLE XXIII

RESIDUAL CORRELATIONS

R_Z-R_Z^+ MATRIX

	1	2	3	4	5	6	7	8	9
33	-15*	05	10	00	-12	-01	13	06	-06
34	-06	-08	05	09	-02	-01	02	05	-05
35	09	-12	04	-01	-05	-07	12	08	-08
36	02	05	-01	-06	05	-03	-02	03	-03
37	11	-01	-07	-04	02	-11	09	08	-08
38	-08	-04	02	10	-10	08	03	04	-04
39	12	-13	-08	09	-21	14	07	-02	01
40	04	11	-08	-07	-13	10	03	11	-11
41	01	01	04	-07	-03	09	-06	-05	05
42	03	-07	01	04	06	-04	-02	04	-04

*Decimal points omitted

TABLE XXIV

CORRELATIONS OF EACH ACHIEVEMENT VARIABLE WITH EACH OBLIQUE FACTOR

S_Z MATRIX (STRUCTURE MATRIX)

	I	II	III
33	241*	310	301
34	133	089	218
35	-003	000	-271
36	-056	-020	-226
37	084	098	-060
38	028	-118	-203
39	-043	-044	-120
40	-003	138	-068
41	080	197	263
42	105	162	143

*Decimal points omitted

TABLE XXV

COORDINATES OF EACH ACHIEVEMENT VARIABLE WITH RESPECT TO EACH OBLIQUE FACTOR

P_Z MATRIX (PATTERN MATRIX)

	I	II	III
33	161*	093	215
34	181	-206	321
35	-101	364	-489
36	-140	306	-400
37	006	229	-209
38	072	-026	-198
39	-063	098	-173
40	-153	410	-307
41	027	030	240
42	050	087	079

*Decimal points omitted

TABLE XXVI

ESTIMATES OF THE THREE OBLIQUE FACTORS EXPRESSED IN TERMS OF OBSERVED VALUES

Teacher Number	$\overline{F}_{I}$	$\overline{F}_{II}$	$\overline{F}_{III}$	Teacher Number	$\overline{F}_{I}$	$\overline{F}_{II}$	$\overline{F}_{III}$
1	2.09	1.42	0.83	33	-0.72	0.06	0.64
2	0.74	-0.81	-0.26	34	-0.34	-0.94	-1.23
3	-0.34	-0.52	-0.36	35	0.94	1.07	0.60
4	-1.70	-0.97	0.58	36	0.82	-0.05	0.72
5	-0.35	-0.06	0.29	37	-0.17	0.00	0.15
6	1.41	-0.90	-1.32	38	0.33	0.26	0.11
7	-0.14	0.67	0.04	39	-0.76	-0.43	0.95
8	1.25	0.50	-0.26	40	0.64	0.37	-0.33
9	-0.39	-0.13	-0.16	41	0.61	0.56	0.12
10	-1.41	-0.87	-0.44	42	-0.72	-1.68	-1.71
11	0.12	-0.58	-0.69	43	0.18	-0.04	-0.02
12	0.02	0.53	0.10	44	0.04	1.10	1.36
13	0.45	-0.77	-1.32	45	-0.23	-0.18	-0.46
14	-0.62	0.26	1.08	46	1.13	0.16	-0.81
15	-0.77	-1.45	-0.31	47	-1.19	-0.68	-0.15
16	3.41	1.86	0.87	48	0.86	-0.10	-0.23
17	0.75	0.74	0.68	49	-0.89	-1.72	-1.21
18	-0.06	-0.42	-0.53	50	-0.57	0.79	0.80
19	-0.71	0.46	0.91	51	0.14	0.27	0.67
20	0.58	1.20	0.29	52	-0.57	-1.10	-1.57
21	0.89	-0.03	-0.20	53	-1.11	-1.02	-0.84
22	0.42	-0.17	1.11	54	-0.62	-0.39	0.01
23	0.83	-0.68	-1.07	55	-0.35	0.08	0.14
24	-0.57	1.02	1.20	56	-0.03	1.17	0.23
25	-1.66	-1.84	-1.37	57	-1.01	0.19	0.17
26	0.34	-0.64	-0.45	58	0.19	0.07	0.34
27	-0.13	1.55	0.09	59	-0.40	-0.73	-0.77
28	-0.83	-0.57	-0.70	60	-1.15	-0.74	-0.57
29	1.26	0.98	0.42	61	0.11	1.53	1.23
30	-1.11	-0.92	-0.64	62	-0.42	0.08	0.55
31	-0.85	-0.46	0.06	63	0.26	0.85	0.74
32	1.29	1.37	1.18	64	0.63	0.85	0.87

The transformed factor scores are shown in Table XXVII. A low number indicates a high "composite" rating.

We will consider that a score of 6 or larger (1 S. D. or more above the mean) indicates that a teacher tends to be considered a poor teacher; and that a score of 4 or less (1 S. D. or more below the mean) indicates that a teacher tends to be considered a good teacher.

Using these standards we find that teachers numbered 1, 16, and 32 would be considered below average since for each of these teachers the three factor scores were greater than six. Teachers 25 and 53 had factor scores of 4 or less on each of these factors. They would be considered as above average as teachers. We find four additional teachers whose factor scores were six or larger for factor estimate one, eight in addition to the three listed above whose factor scores were six or larger for factor two, and eight whose factor scores related to factor three are six or greater.

Six teachers, in addition to the two mentioned, can be considered above average as regards factor one; eight teachers have factor scores of 4 or less on factor two; and seven teachers have factor scores of 4 or less related to factor three.

Thus, if the factor scores (estimates of the three factors) are considered as "composite" estimates of teaching success we find that there are eight "good" and seven "poor" teachers as determined by "composite" one.

Ten "good" and eleven "poor" teachers are identified by "composite" two.

In a similar fashion nine "good" and eleven "poor" teachers are identified by "composite" three.

Part Six: The Prediction Study

In this section the results of the multiple-correlational techniques will be discussed.

As indicated in the preceding part the factor scores were considered as "composite" criteria of teaching success.

The correlations of each variable (10 through 42) with each of the three composite criteria are presented in Table XXVIII.

One observes that the correlations of these thirty three variables with estimated factor one in this table are almost identical to the structure values found in the first column of the S_x, S_y, and S_z matrices. The discrepancies between the correlations of Table XXVIII and the structure values in the S_x, S_y, and S_z matrices are larger for factors II and III. These discrepancies no doubt arise due to the linear transformation (Equation 7, part 5) employed to obtain estimates of these factors. The multiple coefficient of correlation for factor I was higher than the R for either factor II or III, and therefore the two sets of correlations related to factor I should show the greatest agreement. A correlation of .25 or larger was required to indicate statistical significance. Using such a standard, it was found that none of the variables correlated significantly with "composite" criterion one. Variables 28, 31, and 33 were significantly correlated with criterion two; while variables 26, 27, and 33 were significantly correlated with criterion three.

A multiple-correlation coefficient for variables 28, 31, and 33 with "composite" criterion two was computed. R was found to be .43. The intercorrelations were as follows: Variable 28 correlated -.48 with variable 31, and .07 with variable 33. Variable 31 correlated .09 with variable 33.

A multiple-correlation coefficient for variables 26, 27, and 33 with "composite" criterion three was also computed. R was found to be .42. The intercorrelations in this case were found to be $r_{26,27} = -.26$; $r_{26,33} = -.04$; and $r_{27,33} = .13$.

In view of the very low multiple R's reported above it appears that the prediction of the factor scores obtained in this study using temperament, personality, and pre-service measures of achievement is not promising. The two multiple R's reported account for only 18% of the total variance of "composite" criteria two and three.

Since none of the measures were significantly related to "composite" criterion one, no attempt was made to predict this derived estimate of teaching success.

One must conclude that accurate prediction of the three obtained "composite" criteria using the measures and the techniques employed in this study is not possible, or sensible.

The results of each of the six parts of this section have been summarized at the end of each part. The final summary and conclusions will be found in the next section.

SECTION IV

SUMMARY AND CONCLUSIONS

Statement of the Problem

THE PURPOSE of this investigation was to develop equations for the prediction of teaching success with a minimum number of variables and maximum stability as suggested by factor analyses of a variety of measures of the temperament, personality, and achievement of teachers. In order to develop a basis for simplified patterns of prediction a factorial study was made of the interrelationships among the several variables. These factors should point

TABLE XXVII

TRANSFORMED FACTOR SCORES

Teacher Number	$\overline{F}_I$	$\overline{F}_{II}$	$\overline{F}_{III}$	Teacher Number	$\overline{F}_I$	$\overline{F}_{II}$	$\overline{F}_{III}$
1	7.21*	6.50*	6.00*	33	4.24	5.07	5.77
2	5.78	4.10	4.69	34	4.64	3.95**	3.52**
3	4.64	4.42	4.57	35	5.99	6.19*	5.72
4	3.20**	3.92**	5.70	36	5.87	4.94	5.87
5	4.63	4.93	5.35	37	4.82	5.00	5.18
6	6.49*	4.00**	3.41**	38	5.35	5.29	5.13
7	4.85	5.75	5.05	39	4.20	4.52	6.14*
8	6.32*	5.56	4.69	40	5.68	5.41	4.60
9	4.59	4.85	4.81	41	5.64	5.62	5.14
10	3.51**	4.03	4.47	42	4.24	3.13**	2.94**
11	5.13	4.35	4.17	43	5.19	4.96	4.98
12	5.02	5.59	5.12	44	5.04	6.23*	6.64*
13	5.48	4.14	3.41**	45	4.76	4.80	4.45
14	4.34	5.29	6.30*	46	6.19*	5.18	4.03
15	4.19	3.38**	4.63	47	3.74**	4.24	4.82
16	8.60*	7.08*	6.05*	48	5.91	4.89	4.72
17	5.79	5.83	5.82	49	4.06	3.08**	3.54**
18	4.94	4.53	4.36	50	4.40	5.88	5.96
19	4.25	5.51	6.09*	51	5.15	5.30	5.81
20	5.61	6.34*	5.35	52	4.40	3.77**	3.11**
21	5.94	4.97	4.76	53	3.83**	3.86**	3.99**
22	5.44	4.81	6.34*	54	4.34	4.56	5.01
23	5.88	4.24	3.71**	55	4.63	5.09	5.17
24	4.40	6.14*	6.44*	56	4.97	6.31*	5.28
25	3.25**	2.95**	3.35**	57	3.93**	5.21	5.20
26	5.36	4.29	4.46	58	5.20	5.08	5.41
27	4.86	6.73*	5.11	59	4.58	4.19	4.07
28	4.12	4.36	4.11	60	3.78**	4.17	4.31
29	6.33*	6.09*	5.51	61	5.12	6.71*	6.48*
30	3.83**	3.97**	4.23	62	4.56	5.09	5.66
31	4.10	4.46	5.07	63	5.27	5.95	5.89
32	6.36*	6.53*	6.42*	64	5.68	5.95	6.05*

* 1 S.D. or more above the mean.
**1 S.D. or more below the mean.

TABLE XXVIII

CORRELATIONS OF VARIABLES TEN THROUGH FORTY-TWO WITH THE "COMPOSITES" CRITERIA (FACTOR ESTIMATES)

Variable Number	$\overline{F}_I$	$\overline{F}_{II}$	$\overline{F}_{III}$	Variable Number	$\overline{F}_I$	$\overline{F}_{II}$	$\overline{F}_{III}$
10	-17*	-03	18	27	-08	16	25**
11	08	15	17	28	-14	-25**	-15
12	09	14	01	29	-17	-07	-15
13	21	18	05	30	-04	-05	-04
14	15	15	08	31	23	35**	21
15	24	17	-02	32	-19	-19	-06
16	07	03	-04	33	22	25**	28**
17	01	07	-01	34	11	09	13
18	-20	11	19	35	00	-04	-19
19	-01	-02	-03	36	-05	-02	-12
20	-02	-2	00	37	09	06	-05
21	06	-02	-11	38	01	-13	-16
22	18	22	23	39	-04	-08	-08
23	10	23	09	40	03	11	12
24	-03	-09	-21	41	08	19	22
25	-09	-02	-01	42	11	17	07
26	-21	-23	-28**				

* Decimal points omitted
**Significant at the 5% level

the way to better estimates of teaching success.

Measures Used in the Analysis

Forty-two variables were used in the analysis. These may be classified into four groups as follows:

Group (a). Nine estimates of teaching success, consisting of:

1. The principal's acceptability rating of the teacher
2. The principal's rating of the teacher (1st year) on the Wisconsin Adaptation of the M-blank

3, 4. A supervisor's rating of the teacher on the Wisconsin Adaptation of the M-blank

5. The principal's M-blank rating of the teacher (2nd year)
6. Teacher ratings obtained from an outside agency
7. Teacher's self-evaluations
8. Peer evaluations
9. Pupil evaluations

Group (b). Measures of temperament. This group was composed of seven variables derived from the Thurstone Temperament Schedule. They were as follows:

10. Active
11. Vigorous
12. Impulsive
13. Dominant
14. Stable (E for emotionally stable)
15. Sociable
16. Reflective

Group (c). Measures of personality. This group was composed of the sixteen variables obtained from the Sixteen Personality Factor Test. They were as follows:

17. Cyclothymia vs. Schizothymia
18. General intelligence vs. mental defect
19. Emotional stability or ego strength vs. general neuroticism
20. Dominance or ascendance vs. submission
21. Surgency vs. desurgency
22. Positive character vs. immature dependent character
23. Adventurous cyclothymia vs. inherent withdrawn schizothymia
24. Emotional sensitivity vs. tough maturity
25. Paranoid schizothymia vs. trustful accessibility
26. Bohemianism vs. practical concernedness
27. Sophistication vs. rough simplicity
28. Worrying suspiciousness vs. calm trustfulness
29. Radicalism vs. conservatism
30. Independent self-sufficiency vs. lack of resolution
31. Will control and character stability
32. Nervous tension

Group (d). Measures of pre-service achievement. These measures were obtained from University files and were as follows:

33, 34. Practice teaching grades in Education 75 and education methods course

35, 36, 37. Grades in Education 73, 74 and lecture section in Education 75, respectively

38. Percentile rank in high school class
39. Percentile rank on American Council on Education Psychological Examination.
40. Freshman-sophomore grade point average
41. Junior-senior grade point average
42. Composite grade for two education methods courses

Design of the Study

Sixty-four teachers engaged in their second year of teaching in Wisconsin high schools were chosen for this study. This group of teachers had graduated from the School of Education of the University of Wisconsin in 1950. They were a subgroup of the total group of 101 (1950) graduates visited during the beginning teacher visitation program in the fall of 1950. The first four estimates of teaching success were obtained during this visitation while the remaining five estimates were secured during the 1951-1952 school year.

All ratings and grades were converted to numerical scores for the analysis. The Pearson product-moment coefficient of correlation was used throughout the study to indicate linear relationships between various sets of scores.

The matrix of intercorrelations of the teaching success estimates was factored using the multiple-group method of factoring. A further study was made to determine if such factors were factors underlying measures of temperament, personality, and achievement. Estimates of these factors were used as "composite" criteria of teaching success in the prediction phase of this study.

Results of the Statistical Analysis

The questions originally asked in Section I will be answered here. A general factor does not account for the intercorrelations of the nine estimates of teaching success. The three oblique common factors which accounted for these intercorrelations may be designated as:

I. Beginning Teacher Rating Scale Factor: Principal's acceptability rating, Principal's M-blank rating (1st year), Supervisor I M-blank rating, and Supervisor II M-blank rating

II. Second-year Rating Scale Factor: Principal's M-blank rating (2nd year), Outside agency rating, and Teacher's self-evaluation

III. Peer-Pupil Response Factor: Peer evaluation, and pupil evaluation

These factors were found to underlie the several temperament, personality, and achievement measures used in this study. The residual correlations in the proper rectangular matrices were small and therefore further factoring did not seem advisable. Additional factors might have been extracted, but these probably would have accounted for very little of the remaining variance shared by these measures and the estimates of teaching success.

Factor estimates of the three factors used as "composite" criteria would identify eight "good" and seven "poor" teachers (factor I); ten "good" and eleven "poor" teachers (factor II); and nine "good" and eleven "poor" teachers (factor III).

Prediction of these "composites" using temperament, personality, and achievement measures was not successful since the correlations of these variables with the "composites" were exceedingly low, ranging from -.28 to .35. Only nineteen of these ninety-nine correlations were .20 or larger. Correlations of these three types of personal data with the nine original estimates of teaching success were also generally low ranging from -.32 to .38. Fifty-four of these 297 correlations were .20 or larger.

Conclusions

The low correlations of the several temperament, personality, and achievement variables, as here measured, with the nine estimates of teaching success and the three "composites" seems to indicate that the relationship of these measures to teaching success as here measured has not been definitely established.

Although low, thirty-three of the seventy-three more substantial correlations (.20 or larger) reported in this study were correlations of the nine original and the three "composite" estimates of teaching success with the G, M, and Q_3 scales of the 16 P. F. Test and with the practice teaching grade in Education 75. Each of the twelve correlations involving M (Bohemianism vs. practical concernedness) is negative, while the thirty-six coefficients involving G (positive character vs. immature dependent character), Q_3 (will control and character stability), and practice teaching are positive. Due to the consistency of these correlations, as regards sign and magnitude, it would appear that these four measures are slightly related to teaching success.

Since a general factor did not account for the intercorrelations of the estimates of teaching success, one must conclude that these several ratings are not measuring the same "factors." Therefore, the use of a criterion obtained by averaging these nine measures would not be justified. The three derived "composites" of teaching success could be used separately in a program of differential prediction.

BIBLIOGRAPHY

1. Albert, H. R. "An Analysis of Teacher Rating by Pupils in San Antonio, Texas," Educational Administration and Supervision, XXVII (April 1941), pp. 267-74.
2. Allport, Gordon N. Personality, A Psychological Interpretation (New York: Henry Holt and Co., 1937), 588 pp.
3. Bach, Jacob O. "Practice Teaching Success in Relation to Other Measures of Teaching Ability," Journal of Experimental Education, XXI (September 1952), pp. 57-81.
3. Barr, A. S. Characteristic Differences in the Teaching Performance of Good and Poor Teachers of the Social Studies (Bloomington, Ill.: Public School Publishing Co., 1929), 127 pp.
5. ———. "Measurement and Prediction of Teaching Success," Review of Educational Research, XIX (June 1949), pp. 185-90.
6. ——— "The Measurement and Prediction of Teaching Efficiency: A Summary of Investigations," Journal of Experimental Education, XVI (June 1948), pp. 203-83.
7. ———"The Measurement of Teaching Ability," Journal of Educational Research, XXVIII (April 1935), pp. 561-69.
8. ——— "The Measurement of Teacher Characteristics and Prediction of Teaching Efficiency," Review of Educational Research, XXII (June 1952), pp. 169-74.
9. ——— and Emans, Lester M. "What Qualities are Prerequisite to Success?" Nation's Schools, VI (September 1930), pp. 60-4.
10. Baxter, B. Teacher-Pupil Relationships (New York: Macmillan Co., 1950).
11. Bird, G. E. "Pupils' Estimates of Teachers," Journal of Educational Psychology, VIII (1917), pp. 35-40.
12. Blair, Glenn Myers. "Personality Adjustments of Teachers as Measured by the Multiple Choice Rorschach Test," Journal of Educational Research, XXXIX

(May 1946), pp. 652-57.

13. Bossing, Nelson L. Teacher-Aptitude Tests and Teacher Selection, Research in Higher Education, Bulletin No. 12 (Washington, D. C.: U. S. Office of Education, 1932), pp. 117-33.
14. Brandt, Willard J. "A Follow-Up of Some Earlier Wisconsin Studies of Teaching Ability," Journal of Experimental Education, XVIII (September 1949), pp. 1-30.
15. Brookover, W. B. "Person-Person Interaction Between Teachers and Pupils and Teaching Effectiveness," Journal of Educational Research, XXXIV (1940), pp. 272-87.
16. ——— "The Relation of Social Factors to Teaching Ability," Journal of Experimental Education, XIII (1945), pp. 191-205.
17. Broom, M. E. and Ault, J. W. "How May We Measure Teaching Success?" Educational Administration and Supervision, XVIII (April 1932), pp. 250-56.
18. Bryan, R. C. "Evaluation of Student Reactions to Teaching Procedures," Education Digest, XI (December 1945), pp. 55-6.
19. ——— Pupil Rating of Secondary School Teachers (New York: Bureau of Publications, Teachers College, Columbia University, 1937).
20. Buckingham, B. R. "An Unrecognized Ability in Teaching," Journal of Educational Research, VII (1923), pp. 59-62.
21. Buellesfield, Henry. "Causes of Failures Among Teachers," Educational Administration and Supervision, I (1915), pp. 439-52.
22. Burt, Cyril, and Watson, Harriet. "Factor Analysis of Assessments for a Single Person," British Journal of Psychology, Statistical Section, IV (November 1951), pp. 179-91.
23. Butsch, Russell L. C. "The Two-Factor Theory Applied to Measurement of Teacher Traits," Educational Administration and Supervision, XVIII (April 1932), pp. 257-76.
24. Cattell, Raymond B. "A Factorization of Tests of Personality Source Traits," British Journal of Psychology Statistical Section, IV (November 1951), pp. 165-78.
25. ——— "A Note on Correlation Clusters and Cluster Search Methods," Psychometrika, IX (September 1944), pp. 169-84.
26. ——— Description and Measurement of Personality (New York: World Book Co., 1946), 602 pp.
27. ——— Guide to Principles and Techniques of Personality and Ability Testing (Champaign, Ill.: Institute for Personality and Ability Testing), 19 pp.
28. ———, and others. Handbook for the Sixteen Personality Factor Questionnaire: The 16 P. F. Test, Forms A and B (Champaign, Ill.: Institute for Personality and Ability Testing), 18 pp.
29. ——— Personality (New York: McGraw-Hill Book Co., 1950), 689 pp.
30. ——— Personality Structure and Personality Measurement, Proceedings: 1951 Invitational Conference on Testing Problems (Princeton, N. J.: Educational Testing Service, 1952), pp. 82-88.
31. ——— "The Assessment of Teaching Ability," British Journal of Educational Psychology, I (1931), pp. 48-72.
32. Cook, Walter W. and Leeds, Carroll H. "Measuring the Teaching Personality," Educational and Psychological Measurement, VII (Autumn 1947), pp. 399-410.
33. Crane, A. R. "The Changing Role of the Teacher," Forum of Education, IX (April 1951), pp. 113-16.
34. Cronbach, Lee J. Essentials of Psychological Testing (New York: Harper and Bros., 1949), 475 pp.
35. Domas, Simeon J. and Tiedeman, David V. "Teacher Competence: An Annotated Bibliography," Journal of Experimental Education, XIX (December 1950), pp. 101-218.
36. Ebel, Robert L. "Estimation of the Reliability of Ratings," Psychometrika, XVI (December 1951), pp. 407-24.
37. Evans, Kathleen M. "A Critical Survey of Methods of Assessing Teaching Ability," British Journal of Educational Psychology, XXI (June 1951), pp. 89-95.
38. Eysenck, H. J. "Personality," Annual Review of Psychology, III (1952), pp. 151-74.
39. ——— "Uses and Abuses of Factor Analysis," Applied Statistics, I (March 1952), pp. 45-9.
40. Gotham, R. E. "Personality and Teaching Efficiency," Journal of Experimental Education, XIV (December 1945), pp. 157-65.
41. Greene, James, and Staton, T. F. "Predictive Value of Various Tests of Emotionality and Adjustment in a Guidance Program for Prospective Teachers," Journal of Educational Research, XXXII (May 1939), pp. 653-59.
42. Guilford, J. P. "When Not to Factor Analyze," Psychological Bulletin, XLIX (January 1952), pp. 26-37.
43. Haggard, William W. "Some Freshmen Describe the Desirable College Teacher," School and Society, LVIII (September 1943), pp. 238-40.
44. Hampton, Nellie D. "An Analysis of Super-

visory Ratings of Elementary Teachers Graduated from Iowa State Teachers College," Journal of Experimental Education, XX (December 1951), pp. 179-215.

45. Harman, Harry H. "On the Rectilinear Prediction of Oblique Factors," Psychometrika, VI (February 1941), pp. 29-35.
46. Harris, Chester W. and Knoell, Dorothy M. "The Oblique Solution in Factor Analysis," Journal of Educational Psychology, XXXIX (November 1948), pp. 385-403.
47. Harris, Chester W. and Schmid, John, Jr. "Further Application of the Principles of Direct Rotation in Factor Analysis," Journal of Experimental Education, XVIII (March 1950), pp. 175-93.
48. ——— "Direct Rotation to Primary Structure," Journal of Educational Psychology, XXXIX (December 1948), pp. 449-68.
49. ——— "Projections of Three Types of Factor Pattern," Journal of Experimental Education, XVII (March 1949), pp. 335-45.
50. Harsh, Charles M. and Schrickel, H. G. Personality Development and Assessment (New York: Ronald Press, 1950), 518 pp.
51. Hart, Frank W. Teachers and Teaching (New York: Macmillan Co., 1934).
52. Hearn, Arthur C. "Case Studies of Unsuccessful Teachers," Educational Administration and Supervision, XXXVIII (March 1952), pp. 171-4.
53. Hellfritsch, A. G. "A Factor Analysis of Teacher Abilities," Journal of Experimental Education, XIV (December 1945), pp. 166-99.
54. Holzinger, Karl J. "A Simple Method of Factor Analysis," Psychometrika, IX (December 1944), pp. 257-61.
55. ———, and Harman, Harry H. Factor Analysis: A Synthesis of Factorial Methods (Chicago: University of Chicago Press, 1941), 417 pp.
56. ——— "Factoring Test Scores and Implications for the Method of Averages," Psychometrika, IX (September 1944), pp. 155-67.
57. Hult, Esther. "A Study of Achievement in Educational Psychology," Journal of Experimental Education, XIII (June 1945), pp. 174-90.
58. Humphreys, Lloyd G. "Individual Differences," Annual Review of Psychology, III (1952), pp. 131-50.
59. Jersild, Arthur T. "Characteristic Differences of Teachers Who Are 'Liked Best' and 'Disliked' Most," Journal of Experimental Education, IX (December 1940), pp. 139-51.
60. Jones, Ronald D. "The Prediction of Teaching Efficiency from Objective Measures," Journal of Experimental Education, XV (September 1946), pp. 85-99.
61. Kline, Frances F. "Satisfactions and Annoyances in Teaching," Journal of Experimental Education, XVIII (December 1949), pp. 77-89.
62. LaDuke, C. V. "The Measurement of Teaching Ability: Study No. 3," Journal of Experimental Education, XIV (September 1945), pp. 75-100.
63. Lamke, Tom Arthur. "Personality and Teaching Success," Journal of Experimental Education, XX (December 1951), pp. 217-57.
64. Laycock, Sam R. "The Bernreuter Personality Inventory in the Selection of Teachers," Educational Administration and Supervision, XX (January 1934), pp. 59-63.
65. Lins, Leo J. "The Prediction of Teaching Efficiency," Journal of Experimental Education, XV (September 1946), pp. 2-60.
66. Littler, Sherman. "Why Teachers Fail," Home and School Education, XXXIII (1914), p. 255.
67. Lubin, A. and Summerfield, A. "A Square Root Method of Selecting a Minimun Set of Variables in Multiple Regression: II. A Worked Example," Psychometrika, XVI (December 1951), pp. 425-37.
68. McNemar, Quinn. "On the Number of Factors," Psychometrika, VII (March 1942) pp. 9-18.
69. ——— "Statistical Theory and Research Design," Annual Review of Psychology, III (1952), pp. 409-18.
70. ——— "The Factors in Factoring Behavior," Psychometrika, XVI (December 1951), pp. 353-59.
71. Madsen, I. N. "The Prediction of Teaching Success," Educational Administration and Supervision, XIII (January 1927), pp. 39-47.
72. Martindale, Frank E. "Situational Factors in Teacher Placement and Success," Journal of Experimental Education, XX (December 1951), pp. 121-77.
73. Nemec, Lois Gadd. "Relationships Between Teacher Certification and Education in Wisconsin: A Study of Their Effects on Beginning Teachers," Journal of Experimental Education, XV (September 1946), pp. 101
74. Olson, Willard C. Child Development (Boston: D. C. Heath and Co., 1949), 417 pp.
75. Peck, Leigh A. "A Study of the Adjustment Difficulties of a Group of Women Teachers," Journal of Educational Psychology, XXVII (September 1936), pp. 401-16.
76. Pittenger, B. F. "Problems of Teacher Measurement," Journal of Educational

Psychology, VIII (1917), pp. 103-110.

77. Pule, W. H. "The Relation Between Intelligence and Teaching Success," Educational Administration and Supervision, XIV (April 1928), pp. 257-67.
78. Retan, George A. "Emotional Instability and Teaching Success," Journal of Educational Research, XXXVII (October 1943), pp. 135-41.
79. Ringness, Thomas A. "Relationships Between Certain Attitudes Towards Teaching and Teaching Success," Journal of Experimental Education, XXI (September 1952), pp. 1-56.
80. Rolfe, J. F. "The Measurement of Teaching Ability: Study No. 2," Journal of Experimental Education, XIV (September 1945), pp. 52-74.
81. Rostker, L. E. The Measurement of Teaching Ability: Study No. 1," Journal of Experimental Education, XIV (September 1945), pp. 6-51.
82. Ryans, David G. "The Criteria of Teaching Effectiveness," Journal of Educational Research, XLII (May 1949), pp. 690-99.
83. Schmid, John Jr. "Factor Analyses of Prospective Teachers' Differences," Journal of Experimental Education, XVIII (June 1950), pp. 287-319.
84. Schwartz, Anthony N. "A Study of the Discriminating Efficiency of Certain Tests of the Primary Source Personality Traits of Teachers," Journal of Experimental Education, XIX (September 1950), pp. 63-93.
85. Seagoe, May V. "Prediction of In-Service Success in Teaching," Journal of Educational Research, XXXIX (May 1946), pp. 658-663.
86. Simon, D. L. "Personal Reasons for the Dismissal of Teachers in Smaller Schools," Journal of Educational Research, XXIX (April 1936), pp. 585-88.
87. Smalzried, N. T. and Remmers, H. H. "A Factor Analysis of the Purdue Rating Scale for Instructors," Journal of Educational Psychology, XXXIV (September 1943), pp. 363-67.
88. Symonds, Percival M. "Personality of the Teacher," Journal of Educational Research, XL (May 1947), pp. 652-61.
89. Thurstone, L. L. "A Multiple Group Method of Factoring the Correlation Matrix," Psychometrika, X (June 1945), pp. 73-8.
90. ——— Examiner Manual for the Thurstone Temperament Schedule (Chicago: Science Research Associates, Inc., 1950), pp. 11.
91. ——— Multiple-Factor Analysis (Chicago: University of Chicago Press, 1947), pp. 535.
92. U. N. E. S. C. O. The Education and Training of Teachers Toward World Understanding.
93. VonHaden, Herbert I. "An Evaluation of Certain Types of Personal Data Employed in the Prediction of Teaching Efficiency," Journal of Experimental Education, XV (September 1946), pp. 61-84.
94. Wilson, Edwin B. and Worcester, Jane. "Note on Factor Analysis," Psychometrika, IV (June 1939), pp. 133-48.
95. Wisconsin Official School Directory, 1951-1952 (Madison, Wis.: Wisconsin State Department of Public Instruction, 1952), 211 pp.
96. Witty, Paul A. "Analysis of the Personality Traits of the Effective Teacher," Journal of Educational Research, XL (May 1947), pp. 662-71.
97. Woellner, Robert C. "Evaluation of Apprentice Teachers," School Review, XLIX (April 1941), pp. 267-71.

A STUDY OF CERTAIN CRITERIA OF TEACHING EFFECTIVENESS

HAROLD MILTON ANDERSON
University of Colorado
Boulder, Colorado

SECTION I

THE PROBLEM

IN 1915 BOYCE (14) pointed out the need for objective measures of teaching effectiveness. He stated that this need is expressed in three areas, namely: (1) in the guidance of prospective teachers; (2) in the improvement of teachers in service; and (3) in promotion and dismissal. Boyce was concerned with the evaluation of teacher efficiency. As the work in this area was extended, increased concern arose over the criteria of teaching effectiveness. Eighteen years later G. L. Betts (10:99) stated:

> One of the crucial problems in evaluating the education of teachers, if not indeed the crucial problem, is the discovery of a satisfactory criterion of teaching success. With such a criterion at hand, it would be possible immediately to make a scientific evaluation of the worth of various programs for the pre-service education of teachers through a study of the teaching efficiency of large groups of teachers prepared in accordance with those programs. It would be possible to determine what knowledges, skills, and attitudes are most commonly associated with good teaching, and thus to plan curricula definitely directed toward increasing these desirable assets. It would be possible to discover what capacities and personal traits are typical of good teachers, and thus to select for admission to institutions for the education of teachers the candidates most likely to become successful teachers. Scores of other problems relative to such matters as tenure, salary, methods of supervision, the value of practice teaching, age qualifications, requirements for certification, and the like, could be studied experimentally. In the absence of a valid and reliable criterion for teaching success, solutions for these urgent problems must be sought in the realm of expert opinion.

As late as 1949 Ryans (58:696) wrote:

> Certainly until we are able to establish adequate criteria of teaching competency, our whole system of teacher training, appointment, promotion, and tenure fundamentally is on shaky ground.

A multitude of attempts to establish a satisfactory criterion have been made during recent years. The criteria that are in general use at present may be divided in three categories: (1) merit ratings; (2) measures of pupil status and change; and (3) tests of abilities thought essential to success. Rating scales may be further classified with respect to the person using the scale, namely the supervisor, administrator, pupil, teaching colleague, or the teacher himself. Pupil status and change is measured by some form of test which is designed to estimate the extent to which certain objectives of instruction are attained. In some cases these tests cover only certain selected aspects of subject matter, but often they attempt to discover status and change in areas such as attitudes, interests, appreciations and information. Evaluations of this type are sometimes made for a single school, city, county, or even an entire state. Teacher examinations usually measure professional and subject matter information, mental capacities, cultural status and professional interests and activities.

This study is concerned with the degree to which the various criteria of type 1 and 2 agree in their rankings of teachers.

Statement of the Problem

The primary purpose of the study reported herein is to investigate the relationships among certain criteria of teacher effectiveness by the use of comparative and correlational techniques. A secondary aim is to trace the historical development of these criteria and to note justifications offered for their use. The investigation also serves as a followup of selected teachers holding the University of Wisconsin Teacher's Certificate.

It should be born in mind that this is not an

---

* The author wishes to express his appreciation to Dr. A. S. Barr, his major professor, for his encouragement, direction, and inspiration throughout this investigation.

attempt to validate criteria of teaching competence. Such validation implies that a known standard exists and that comparisons can be made to discover the extent to which the particular criterion correlates with the standards. In this study no such standard is accepted or postulated.

Specifically the investigation seeks to answer the following questions:

1. What justifications are commonly offered for the use of various criteria?
2. To what extent are these criteria correlated?
3. To what extent do the different estimates made by the use of the same instrument on the same group of teachers by different raters agree?

SECTION II

AN OVERVIEW OF THE CRITERIA SELECTED FOR STUDY

THE EDUCATIONAL literature is filled with materials relative to the criteria of teacher effectiveness. Some of the material is based on research; some on philosophic principles and logic; and some primarily on opinion, often founded on casual observation and intuition. The reader needs to exercise considerable judgment and discretion in picking his way through the relevant and irrelevant.

The following discussion represents the material that the writer felt significant in considering the strengths and weaknesses of the selected criteria. The organizational plan takes up first rating system in general, then proceeds to specific rating categories designated by the classification of the rater, continues with a brief consideration of teacher evaluation by pupil achievement, and concludes with a brief summary.

If the criteria of teaching effectiveness are to be considered in their proper perspective, cognizance must be taken of the foundations upon which their validity rest. Every criterion is an arbitrary standard which serves to evaluate. Most of the criteria in common use have been validated by some accepted system of values. Cook (in 43:1) states the case very well as follows: "The value of measurement depends on the extent to which the relationships established are crucial from the social point of view." The ultimate validation of criteria of teaching effectiveness rests with its agreement with the composite judgments of the individuals who are concerned about the problem.

Pittenger (50) states that logically there are three bases for estimating teaching success or ability. It is possible to judge effectiveness by the results produced, by the processes employed in teaching, or by the equipment that the teacher possesses for teaching. The last two items will be considered along with rating scales while the first will be considered with pupil achievement.

Rating Scales as Criteria

It seems inevitable that teachers will be evaluated with or without the aid of a measuring device or instrument. If such evaluations can be made to agree with the judgments of society to a higher degree of consistency through the use of some instrument, then the use of such a tool is valid for that purpose. A number of rating scales have been developed with that objective in mind.

The rating scales that have been widely used in the field of teacher evaluation may be separated into two broad categories: (1) general ratings which are all-inclusive regarding the merit of the teacher; and (2) specialized scales which consider traits and qualities or note specific behaviors which are believed to reflect the teaching ability of the teacher.

The first of these types is criticized in that it is a subjective personal evaluation which may fail to give proper consideration and weighting to all the important factors contributing to teaching success or failure. A rating or ranking is given which states the position held by the teacher with reference to other teachers according to the standards of the rater.

The second is often criticized with respect to the method of selecting and weighting the factors used. Experts in the field often disagree as to which traits, behaviors, and qualities are essential to teaching success and the extent to which each contributes. Such a scale does serve to emphasize those traits, qualities and behaviors which the scale producers have selected as the attributes of good teaching. When a scale of this type is used with a number of raters, it may be considered as the definition of a good teacher. A number of subjective judgments have been substituted for the grand comprehensive evaluation of the first type. There is a strong tendency for the rater who has a strong conviction regarding the overall merit of the ratee to rate him accordingly on all items of the point scale. This tendency is known as the "halo effect."

The semantic problems in scale construction are great. The use of poorly defined terms tend to reduce the inter-rater objectivity of rating scales. The words employed often have different meanings for different raters. Ambiguity must be avoided. Detailed definitions of terms and personal conferences among raters might be very helpful in this respect.

One of the outstanding weaknesses of rating scale procedures is the tendency of the rater to over-rate the ratee. An excellent example of this is the in-service follow-up by Goetsch (28) of 237 Iowa State Teachers College graduates by means of a rating scale incorporating 20 characteristics. He found that the teachers were rated as follows: superior, 30 percent; average, 15 percent; above average, 52 percent; and below average 3 percent. While the writer has only the highest regard for the training institution involved, he does seriously question that only eighteen percent are only average or less. Knight (40), in studying this tendency to over-rate, found that it increased with the length of acquaintance of rater and ratee. Perhaps the best explanation of the tendency to over-rate lies in the fact that raters substitute their own standards for those employed in the rating device. Symonds (60:693) expresses this very aptly:

> Every supervisor has his own expectations and unique requirements in a teacher. ....It will never be possible to find agreement in teacher rating as long as principals and supervisors have different standards for judging teachers.

Symonds emphasizes the fact that ratings should be made only after considerable time has been spent in observation. The impression obtained in the first few minutes may be deceiving in that the classroom containing an observer is not the natural habitat of the student and his reactions differ accordingly.

The bitterest attacks on rating scales pertain to the lack of reliability in these instruments. The variations in results that occur may come as a consequence of change in the teacher, change in the observer, or in the interpretation of the scale. Ratings are certainly affected by the length and conditions of observation, the attitudes of persons involved, the social and emotional status of the pupils, the prejudices of the observer, and the nature of the occasion when the last rating took place. It is not at all surprising that the statistical reliabilities quoted in the literature show such large variations.

Frequent reference is made to a statement by Rugg (57) that the reliability of a rating varies directly as the square of the number of observers rating independently. While this article was written some years ago, it still has something to consider for those interested in ratings. It discusses the requirements and conditions under which the increase in reliability may be expected. The statement should be considered only in conjunction with the conditions described in the article.

Schultz (59) warns that effectiveness can not be accurately estimated by a summation of rating which is essentially the same as the averaging process. He makes a point of stressing the fact that evaluative processes should include the person being rated and should be for the purpose of improving teaching.

In spite of the deficiencies of rating scales, they continue to be used. One needs to remember, as Hansen (33) points out, that teachers are often hired or fired on the basis of subjective information only. The practical world insists that decisions be made. Certainly a valid evaluation might be exceedingly helpful for the guidance, protection, and improvement of teachers.

Baier (3), in discussing officer rating systems states: "In some cases ratings may be the best criteria because value judgments are essential elements." The same may be true of teacher evaluation.

Specific Types of Ratings

Teacher ratings may be made by the teacher himself, by supervisors, by pupils, by peers, and others. Perhaps the most frequent type of rating is that by the supervisor. One of the outstanding advantages of this type of rating is that a number of teachers within the system may be ranked in order of excellence according to the supervisor's standard. Inasmuch as the supervisor has the opportunity to observe the teacher on more than one occasion, the probability of that judgment being made on an unusual performance is very slight. Since supervisors are expected to diagnose difficulties and aid teachers as part of their duties, they may be expected in general to be well qualified to judge teaching competence. At the times when ratings must be obtained for areas greater than that served by a single supervisor, the ratings are subject to additional error arising from differences in raters.

Teachers have also been evaluated by students. The basic assumption is that because of his close contact with the teacher the pupil has information not possessed by others for rendering sound judgments about the effectiveness and ability of the teacher. The use of such rating is questioned by many, however, who argue that pupils lack the maturity to judge competently. Lack of experience with teachers in any quantity prevents the differentiation of superior and mediocre teachers on the lower grade levels. It is often further argued that they lack the experience and maturity to pass judgment on the value of the course contents. Nonetheless, Cook and Leeds (21), and Albert (2) state that pupil evaluation of teaching is reliable, valid, and practical. Bryan (18) urges the use of pupil evaluation because: (1) the learner's attitude is significant in the learning situation; (2) ratings emphasize the need of improvement; (3) an increase in concern over teaching problems by the staff; (4) pupil ratings are an index of effectiveness because teacher influence is proportional to the respect and admiration of pupil for instructor; and (5) trains pupils to be critical citizens in a democracy. Boyce and Bryan (15) found that pupil opinions tended to be quite stable over a five-year period. Some universities (37, 51) use student ratings for faculty evaluation, both for the purpose of improving instruction and reemployment.

Ratings of instruction by colleagues, while not as popular as pupil evaluation, has been given considerable support. It is argued that by interclass visitations and through the school activities, the peers become sufficiently familiar with the work of the other teachers to pass judgment on the competency of their teaching. It is further argued that the process is democratic and important in that respect. Evaluation by peers presents problems, particularly if no means are provided for interclass visitation.

Many persons have advocated a system of self-evaluation, usually in order to improve in service. It would seem that the teacher is in the best position to judge the quality of work which he is doing. A suitable standard for comparison is missing in most cases. Seldom, if ever, can a person evaluate his own work objectively.

It can be readily seen that we have a number of criteria available. Each of these criteria has at least one difficulty associated with it as well as some advantages. Cureton (in 43:639) states the case for rating scales very well:

> When well constructed and rightly used, behavior check lists and rating scales help materially in making the post recording and judgments of different observers more comparable, in reducing differences in the interpretation of the basic definitions and agreements, and in minimizing the disturbing effects of selective attention and selective recall.

Pupil Achievement as a Criterion

Teacher evaluation experts are almost universally agreed that the measure of true effectiveness as a teacher is the change that is produced in the pupils taught by that teacher. While one may agree that pupil gain is the only valid criterion of teaching effectiveness, validity implies that such a criterion also be reliable as well as relevant. Reliabilities reported in pupil gain studies have been consistently low.

Many school people use the criterion of pupil status to evaluate instruction. Some school districts require that pupils take standardized tests at the completion of the school year. Certain states have statewide testing programs and pupil promotion is based on the score produced on that test. The teacher is evaluated on the basis of whether her pupils pass or fail. While this practice of teacher evaluation is decreasing, it is not extinct.

Such a practice is unfair in that it does not take into consideration the ability of the pupils to learn nor the conditions under which the learning takes place. An improvement was made when schools came to use the achievement quotient which considered the intelligence of the pupil. Using the formula score equal to the general average of the class plus one-hundred minus the average I. Q. of the class, Brooks (17) rated the teachers in his district. He says,

> In addition to a substantial general raise in salary throughout the district, the school boards were persuaded to grant special increases of one or two dollars per week to teachers who rated ninety-five percent or better.

With the passage of time these procedures fell into disrepute with both curriculum experts and statisticians.

The modern educationist claims that schools must teach more than subject matter and that attitudes, appreciations, ideals, and personality improvement are equally important.

Coy (23) in 1930 warned that the accomplishment quotient could be used to evaluate teachers only under special conditions. Modern educators realize that many other factors besides intelligence affect pupil grades and have used regression equations and similar techniques in an attempt to predict the scores that pupils should obtain if instruction were equal. The discrepancies between the predicted scores and the actual scores have then been used as a measure of the effectiveness of the teaching.

Almost any book on measurement (54, 43) will discuss the measurement errors to which our test procedures are subject. The test scores of a pupil simply indicate that, in all probability, his "true" score is within certain limits of the score he produced on the test. Also, pupil accomplishment is affected by many factors which we are unable to measure satisfactorily at the present time and it is questionable if we ever will. Many agencies other than the teachers affect learning. Parents, books, playmates, television, radio, and many other forces motivate and direct learning in present day society.

A common danger which arises out of the testing of subject-matter is the tendency of teachers to prepare for the test by urging the student to cram and memorize test items rather than emphasize understanding and broad concepts.

A great deal of emphasis has been placed on pupil gain as a measure of teaching effectiveness. Thorndike (in 43) points out that difference scores have a low reliability, often approaching zero. Thus, they may have little or no validity as measures of teaching effectiveness.

Many tests are so constructed that a large gain is quite easily obtainable if the pupil has a low pre-test score, but gain may be restricted if the original score was high. This is especially true if the test has a low "ceiling". Thus, if pupil gain is considered as the difference between the pre-test and the final, the pupil who had a high score on the pre-test is under a considerable handicap. Special statistical and graphical techniques have been used to compensate for this handicap. In most instances they have not proved to be entirely satisfactory.

Summary

A study of the criteria used for the evaluation of teaching effectiveness leads one to appreciate the difficulties present in all phases of such research. The importance of evaluating teachers has forced educational leaders to use such methods as exist since new methods have not been developed. The great questions facing education today in this area are the following:

1. Can these criteria be used singularly and independently or must composite measures be formed?
2. If composite measures are formed, what weighting must be given each component?
3. Can other measuring devices of greater reliability and validity be developed?

SECTION III

THE METHOD OF THE INVESTIGATION

The Design of the Investigation

DURING THE fall semester of the 1950-1951 school year, 101 graduates of the College of Education, University of Wisconsin, who hold the University Teachers Certificate were visited while teaching at the school where they were employed. These visits were part of a followup of its graduates by the School of Education. The purpose of the visitation program was two-fold. In the first place, an attempt was made to gain information concerning the problems of these beginning teachers and given them any help possible. In the second place, an evaluation of their teaching was made in order to check the results of the University program of teacher training.

The success of these new teachers during their first semester was determined by several methods. A team of two visiting supervisors rated them on a short form of the Wisconsin Adaptation of the M-Blank during their trip to the school. At the same time they secured an "acceptability rating" from the principal by means of an interview. The "acceptability criterion" was broader than the usual evaluation in that it was concerned with the teacher as a member of both the school and the community. Another measure of the beginning teacher's effectiveness was obtained on a "participator's rating scale".

After considering the criteria employed in the 1950-1951 studies, the writer felt there was a need to see if the criteria used were comprehensive enough to measure teaching effectiveness adequately. This is not intended to be a criticism of the previous study which was very well done by thoroughly competent and sincere individuals. The previous investigations served to stimulate this writer to search the literature for a definite relationship among the criteria commonly employed in evaluating teaching effectiveness. Specifically stated, the problem settled upon was to discover what relationships, if any, that exist among such criteria as principal's ratings, pupil ratings, peer ratings, self rating, and criteria based on pupil achievement.

The writer worked with the Teacher Personnel Research Committee, University of Wisconsin. The Committee gathered considerable data about Wisconsin graduates in the field during 1951-1952.

One descision of the committee was to set up a testing program for the pupils taught by the teachers to measure their attainments in the subject matter areas. The original plan called for a measurement of achievement in the area of concommitant learnings as well as foundational learnings. After some discussion the measurement of concommitant learnings was omitted by the committee for two reasons. They felt that such items as appreciations, attitudes, interests, and personality are subject to large measurement errors. Grave doubts were expressed as to our ability to measure gains in these areas. Furthermore, since the pupils taught by these teachers are in high school and subject to the influences of a number of teachers as well as classmates and others, it was decided that if gains were detected in these areas, there would be no real justification for ascribing these gains to the teacher being studied. Since the subject matter areas tend to be more specialized, such changes might reasonably be expected to be the effect produced by the teacher in that area. It was recognized that some changes may result from maturation. The comparison is, however, of teachers whose classes contain a number of pupils, where in it seems reasonable to assume that the mean of the gains resulting from maturation will be approximately the same for the various schools. The addition of a constant to the mean gain score produced by the teacher is immaterial because the purpose is to rank the teacher rather than determine his absolute effectiveness.

In consideration of the fact that factors other than the teachers influence learning, measures were obtained of the status of the pupils with respect to intelligence, age, and average scholastic standing in order to determine how much of the learning could be attributed to these factors, assuming that the effects of all other factors are negligible. After the relationships of the selected factors to pupil change had been determined, regression equations were set up for predicting the gain to be expected. The mean discrepancy for each class between the actual scores and the predicted scores may then be considered a measure of teaching effectiveness. Such a measure is sound only if the variations that are due to differences in the groups of those variables which affect learning have been removed.

Since the scores of pupils on achievement examinations are occasionally regarded as a criterion of teacher efficiency, it was decided to use this as one of the criteria to be studied. Accordingly, a score was predicted for each pupil by means of a regression equation. The mean difference for each class between predicted scores and attained scores were used as a measure of teaching effectiveness. When certain unavoidable delays made it apparent that about three months of school would elapse before the initial testing, it was decided to use both initial and final test discrepancies as measures of teaching ability. It was felt that if a teacher could be found who would place at the top in all three of these pupil achievement evaluations, that teacher could be considered as an excellent one.

The ratings of the teachers by themselves, their peers, their pupils, the principal, and another agency were collected through correspondence. The intercorrelations among these criteria were computed and will be presented later.

Representativeness of the Group Selected

It would be difficult to set up pupil achievement criteria for all the teachers in the group that were visited because there was great diversity in the classes taught. Accordingly, the decision was reached to select a group of 30 to 35 teachers whose classes would be most amenable to the testing program.

Since the study is concerned with criteria of the teaching effectiveness of Wisconsin graduates, it was decided to select the group to be studied from the group visited last year. Of that group only 64 were teaching in Wisconsin high schools during the current year. The remainder had left the teaching field for the armed services, marriage, more lucrative positions, or other fields that seemed to be more compatible with their interests and abilities. The group under consideration was augmented by selection of other 1950 graduates of the University of Wisconsin teaching in the state high schools. After a preliminary screening a questionnaire form (Appendix IIa)* was sent to each teacher on the list of eligibles. Approximately 89 percent of the group returned the questionnaire.

The 73 returned questionnaires were examined with respect to courses, length of class periods, class size, grade level, textbook used, and willingness of the teacher to cooperate. On the basis of this inspection, 32 teachers were selected for participation in the study. Later two teachers were dropped from the study. In one case the teacher changed positions during the school year. The other was found to be teaching in a position which differs greatly from the circumstances under which most of the University of Wisconsin graduates teach.

Almost every attempt to show that a sample is representative of the total population from which it is drawn is predestined to failure. The investigator may stratify the population with respect to as many variables as he may desire and ascertain that his sample is representative of this larger population insofar as the characteristics considered are concerned. However, since the number of variables which may be used to compare groups is almost infinite, the investigator must draw the line at some point and assume that further comparisons are unnecessary and irrelevant.

The comparisons are irrelevant when the basic criterion for comparison does not materially affect the trait which is under investigation. Since this study is concerned with the relationships of various criteria of teaching effectiveness, and no information is available at the present time as to the way in which factors affect teaching effectiveness as measured by all these criteria, there is no characteristic which could not be assumed to be irrelevant. However, since it is reasonable to suppose that there are some factors which do affect teaching effectiveness, an attempt is made to show the group representative of graduates of the University of Wisconsin.

The writer is aware that there is some risk involved in making inference from this sample to the larger population. However, he feels that to restrict any conclusions which might be drawn to the group studied alone is pointless. It is pointed out that inferences drawn in sampling studies may also be in error, but possess the advantage that the investigator is in position to state the probability of being in error.

---

* All references to Appendices may be found in the Library, University of Wisconsin.

It may be profitable to see how these selected teachers compare in some respects with another group of 1950 College of Education graduates. Mr. Harley E. Erickson of Superior State College, Superior, Wisconsin, was concurrently conducting a factorial investigation of the teaching abilities of the 64 teachers frequently referred to in this section. Table I gives a brief comparison of the two groups.

The small differences which exist in mean percentile rank on the Psychological Examination and in high school class between the two groups may possibly be explained in that our group is composed of teachers in the academic courses and does not include fields of special abilities such as fine arts, home economics, industrial education, and music. The sex difference may possibly have risen from the same cause.

In Table II the group is compared with groups of the 1948 and 1949 graduating classes of the College of Education, University of Wisconsin, on basis of means and standard deviations in general grade point average. It would appear that the group studied herein is slightly superior to the other groups in this respect. This difference may possibly be the result of selective processes such as the restriction of the group to teachers of academic subjects and that by the second year a number of people have left the teaching profession.

The teachers studied in this investigation may be considered as representative of the population of Wisconsin College of Education graduates teaching in this state. They have met the requirements for entrance to the University and to the College of Education. While enrolled in the College of Education they have all passed through approximately the same training experience and indoctrination. Upon graduation they have secured positions in the state high schools. In these and other respects they have certain patterns of homogeneity that are inescapable. While there are many aspects in which the parent population is heterogeneous this is also true of the Our Group.

Insofar as can be determined at the present time, the group does not differ radically from the parent population. Small differences do exist in the following: (1) the courses taught; (2) slight differences in mean and standard deviations of grade point averages, percentile rank in intelligence, and percentile rank in high school class, (3) expected differences in individuals. Since the same selective processes might be expected to operate in the future as have in the past, one might well expect this group to be truly representative of teachers obtaining the University Teachers Certificate and employment in the State of Wisconsin. Inferences drawn from this sample can be applied to the general population if one wishes to make the assumptions noted above.

Data-Gathering Devices Employed

The measuring devices used in this study may be considered in two groups: (1) those applied to the teacher, and (2) those applied to pupils taught by the teacher. The measures applied to the teachers are rating scales of various types. Three of these scales are adapted from Student Reaction to Instruction Questionnaire developed my Roy C. Bryan of the campus school at Western Michigan College of Education, Kalamazoo, Michigan. The measures applied to students are achievement tests in subject matter areas. The tests used in this study were selected after securing the opinions of experts in these particular subject matter areas, by reference to Buros (66) and by inspection of copies of the tests. No test was used if its reliability did not exceed 0.90. A form (Appendix II. B) was sent to each school to secure information about the age, intelligence, and average scholastic standing for each pupil in the classes where the achievement tests were given.

The teacher rating scales used are listed:

1. A short form of the Wisconsin Adaptation of the M-Blank. Originally M-Blank, Data for Individual Staff Members, 1940 edition, Cooperative Study of Secondary School Standards. (Appendix II. C)

2. A pupil rating form—adapted from Bryan. (Appendix II. F)

3. A teacher self rating form—adapted from Bryan. (Appendix II. E)

4. A peer rating form—adapted from Bryan. (Appendix II. D)

5. Administrative X Criterion (confidential, i.e., not open for inspection).

Pupil Achievement Tests:

1. Typing—Competent Typist Test
   Gregg Awards Department, 330 West 42nd St., New York 36, N. Y. (Appendix III. A)

2. Mathematics 7—Colorado State College of Education Basic Skills in Arithmetic Test, Form A, Grades 6-12 1945 Chicago Science Research Associates. (Appendix III. B)

TABLE I

COMPARISON OF VISITED GROUP AND OUR GROUP

Characteristic	Our Group	Visited Group
Number	30	64
Sex	17 F 13M	30F 34 M
Mean Grade Point	1.96	1.94
Mean Percentile, American Council, Psychological Examination	69.0	65.8
Mean Percentile in High School	84.5	75.7

TABLE II

COMPARISON OF GROUP WITH 1948 AND 1949 GROUPS ON SCHOLARSHIP

Characteristic	Sample	1948	1949
Number	29	205	102
Mean	1.96	1.88	1.75
Standard Deviation	0.42	0.38	0.40

TABLE III

MEANS AND STANDARD DEVIATIONS OF VARIABLES (COURSE I)
(N = 113)

Item	Age	P.R.	S.A.	P.T.	F.	G.
Mean	15.15	39.59	6.27	13.30	17.34	4.03
S. D.	0.613	26.01	2.32	10.69	12.52	6.63

3. English—Smith, D. V. and McCullough, C. M. Essentials of English Test Form A, 1939 Minneapolis Educational Test Bureau. (Appendix III. C)

4. General Science—Educational Testing Service Cooperative Science Test, Form Y 1948 New York Educational Testing Service. (Appendix III. D)

5. Algebra—American Council on Education, Cooperative Algebra Test Form 8 1942 New York Cooperative Test Service. (Appendix III. E)

6. World History—American Council on Education, Cooperative World History Test Form Y, 1948 New York Cooperative Test Service. (Appendix III. F)

Sources and Collection of the Data

The methods employed in collecting the data for various segments varied considerably. It is believed that this exposition of the sources and methods of collecting of data will be presented most lucidly if the data relative to each criterion are presented separately.

The rating of the teacher by his principal was secured on the one page form of the Wisconsin Adaptation of the M-Blank. A sample of the blank and the letter sent with it to the supervising principal are in Appendix II. C. The principal is expected to list what he considers to be the outstanding elements of strength and weakness in the work of the teacher and to make any comments which he deems appropriate. The principal was also asked to give an overall evaluation. In this particular study the general evaluation only is considered and is recorded in column M of Appendix I. B. The ratings run from 1 for superior to 5 for inferior. Teachers and courses are indicated by coded numbers as a precaution lest the rating of a teacher be used to his disadvantage by some unscrupulous individual.

As a result of the fine cooperation of the participating principals, the ratings of the teachers were secured from five pupils taught by each teacher. In several schools the administration felt that pupil ratings were not desirable for their system. These wishes were respected and the appropriate spaces in the data sheet remain blank. Five forms and five stamped addressed envelopes were sent to the principals with the request that five pupils be chosen at random to rate their teacher. The instructions sent out with all rating scales were that the rater remain anonymous. The form consists of 9 questions on which the pupil is asked to rate the teacher. In an attempt to overcome the tendency to overrate, the pupils were asked to divide all the teachers they have had into five equal groups according to excellence. The pupil is then asked to indicate which fifth the present teacher fits. In recording the data in Appendix I. B. columns, 2 to 6 inclusive, 1 indicates the best fifth, 2 for the second fifth and so on until the worst fifth is 5.

The peer rating form (Appendix I. D) is very similar to the pupil rating, except some subscales for rating differ. This form was sent to two teaching colleagues of the teacher. The results are tabulated in Appendix I. B.

A self rating form was also sent each teacher with instructions to send in a rating of his own teaching. The form which was an adaptation of Bryan's Pupil Rating Scale is very similar to the peer and pupil scales. (Appendix II.E) The same scale of values is used in recording the data as in the other two cases. (Appendix I. B)

Since the Administrative X criterion is confidential, no information about it other than the method of scoring will be given here. The largest numerical score that may be attained is five. The higher score indicates the poorer teacher.

There are 7 areas in which pupil achievement was studied: English 9, English 11, Typing, Mathematics 7, Algebra, General Science, and Modern History. A few teachers had classes in two areas and were tested in both. Since achievement in these subject matter areas was to be compared on the basis of standard scores, a minimum of five classes in each area was desired. Several of the course areas were below this minimum. Consequently, several test groups were set up in several cities using teachers who were comparable in training and experience. Data for these groups are located in the Appendix I. A with letters as code for teachers rather than numbers.

The pupil achievement program was carried on with the aid of the school principals. Tests were sent out with a letter (Appendix II. B) to the principal requesting that the test be administered under his supervision. Most of the tests were given in December 1951. The tests were returned to the Teacher Personnel Research Committee which scored them and sent copies of the scores to the teacher and principal involved. In the spring tests were again sent to the principals with instructions to retest the group 70 or 75 school days after the initial testing, depending on the course. So far as can be determined, all tests were given on schedule except one. That test was given 68 days after the first testing instead of the prescribed 70. No adjustment was made for this time discrepancy because it might possibly be used to identify the teacher. The initial test scores together with the measures of intelligence, age, and average scholastic achievement are tabulated by classes in Appendix I. A.

The Treatment of the Data

The intercorrelations of age, intelligence, and average scholastic standing were calculated for each course. These correlations were tabulated in a matrix. Using the Aitkin Method of Pivotal Condensation (62:89), prediction equations were set up for each course to predict the pretest scores. The mean of the discrepancies between the actual scores and the predicted scores was obtained for each class. The class discrepancy was regarded as a measure of the effectiveness of that teacher. A standard score for a teacher was calculated by subtracting the mean discrepancy for the entire course from the mean discrepancy in the class and dividing the difference by the standard deviation of the mean discrepancies for the course. The teacher may be ranked by these standard scores if the assumption is made that the teachers in one course are equally competent as those of another. Rather than make this assumption it was decided to assume that the norms provided with the tests are all equally applicable to their groups. A constant of 5 was added to each standard score, rendering it positive. By multiplying the scores by a positive number all values are positive. The scores of the different groups should be comparable after they have been multiplied by the ratio of the mean score in the group to the mean of the norms for that group. These scores may now be considered to rank the pupils according to the achievements of their pupils after compensation for differences in age, intelligence, and average scholastic standing. An identical procedure was used for the final test score except that final score was substituted for initial score, and that the pretest score was used as a predictive variable. In both cases there were teachers who were teaching classes in two courses tested. The two ratings which the teacher achieved were averaged in determining their ranking.

In working with pupil achievement as a measure of teaching effectiveness, it is common to consider the change produced rather than the status. If the initial score is subtracted from the final score, the difference is the gain score. In this investigation the mean of the gains scored on the various segments of the pretest range were plotted against the segments of the pretest scores for each course. If no appreciable relationship was found, the gain scores were left unchanged. In some of the cases a trend was found for the low pretest scores to be associated with the high gains. To compensate for this, each gain score was multiplied by the ratio of the mean of all the gains to the mean of the gain scores in the segment of pretest scores where the corresponding pretest score is located.

The intercorrelations of the adjusted gains with age, intelligence, scholastic achievement, and pretest scores were used to determine a prediction equation for gains. The mean discrepancy of actual gain minus predicted gain is used as a discrepancy score for each teacher. Teachers were rated on these discrepancy scores by the same process as on the pretest and final score discrepancies.

In some of the cases it was found that the prediction equation accounted for less than 25 percent of the variance. In these cases the actual mean gains were used as measures of the teachers' effectiveness after the mean gains had been converted to rankings by the standard score techniques mentioned before.

After ratings and rankings of the teachers have been obtained by all criteria, the relationships among these were determined by the Pearson product moment correlation formula.

SECTION IV

THE ANALYSIS OF THE DATA

THE MAJOR purpose of this investigation was to study the relationships among certain criteria of teaching effectiveness. Data regarding the success of thirty graduates of the College of Education, University of Wisconsin, in the second year of service were collected. The types of data collected with a short explanation of the symbols employed and the scoring system of each will be reviewed. The original data are presented in Appendix I.

1. A rating of teacher by principal on a one page form of the Wisconsin Adaptation of the M-Blank. Symbol M; scoring range of one to five with the lower score indicative of the better teacher.

2. A rating of the teacher by pupils on an adaptation of the Bryan Reaction to Instruction Questionnaire. Symbol Pu; scoring same as above for M. Total of five pupil ratings used in calculations.

3. A rating of the teacher by himself on a specially adapted self evaluation form. Symbol Se; scoring same as M above.

4. A rating of the teacher by his colleagues on the peer-evaluation form. Symbol Pe; scoring as M above. Sum of two peer ratings used in calculations.

5. A rating of teacher by another agency on a special Administrative Criterion. Symbol AX; scoring range of five upward. Lower score indicates better teacher.

6. Pupil scores on an initial administration of standardized achievement tests. Courses

TABLE IV

INTERCORRELATIONS OF VARIABLES (COURSE I)

	Age	P. R.	S. A.	P. T.	F.	G.
Age	....					
P. R.	- .272	....				
S. A.	0.270	- .585	....			
P. T.	- .138	0.597	- .550	....		
F.	- .223	0.664	- .555	0.848	....	
G.	0.349	0.291	- .169	- .116	0.520	....

TABLE V

PREDICTION EQUATIONS FOR COURSE I

P. T. '	= 0.17 P. R.	-1.40 S. A.	+15.3	
F'	= 0.11 P. R.	- .24 S. A.	+ 0.81 P. T.	+ 3.8
G'	= 5.00 Age	+0.11 P. R.	-76.0	

TABLE VI

MEANS AND STANDARD DEVIATION OF VARIABLES (COURSE II)
(N = 79)

Item	Age	P. R.	S. A.	P. T.	F.	G.
Mean	15.24	51.75	5.29	9.23	15.37	6.15
S. S.	0.875	25.62	2.36	4.75	4.97	4.09

TABLE VII

INTERCORRELATION OF VARIABLES (COURSE II)

	Age	P. R.	S. A.	P. T.	F.	G.
Age	......					
P. R.	- .103	......				
S. A.	- .009	- .352	......			
P. T.	- .342	0.381	- .266	......		
F.	- .169	0.369	- .372	0.615	......	
A. G.	0.056	0.174	- .205	0.003	0.682	......

TABLE VIII

PREDICTION EQUATIONS FOR COURSE II

P. T.' = -1.70 Age +0.054 P. R. - .332 S. A. -34.2

F.' = 0.019 P. R. -0.415 S. A. +0.548 P. T. +11.50

A. G.' not used; cannot account for 10% of variance.

TABLE IX

MEANS AND STANDARD DEVIATIONS OF VARIABLES (COURSE III)
(N= 99)

	Age	I. Q.	S. A.	P. T.	F.	A. G.
Mean	12.40	105.64	5.75	46.13	54.54	8.555
S. D.	0.62	12.90	2.67	10.63	10.12	4.58

TABLE X

INTERCORRELATIONS OF VARIABLES (COURSE III)

	Age	I.Q.	S.A.	P.T.	F.	A.G.
Age	......					
I.Q.	-.583	......				
S.A.	0.361	-.633	......			
P.T.	-.197	0.481	-.790	......		
F.	-.238	0.494	-.704	.870	......	
A.G.	-.074	0.128	-.004	-.074	0.337	......

TABLE XI

PREDICTION EQUATIONS FOR COURSE III

P.T.' = 1.732 Age -3.288 S.A. +43.562

F.' = 0.085 I.Q. +0.095 S.A. + 0.797 P.T. +8.279

A

A.G.' not computed; can only account for about four percent of variance.

TABLE XII

MEANS AND STANDARD DEVIATIONS OF VARIABLES (COURSE IV)

(N = 95)

	Age	P.R.	S.A.	P.T.	F.	G.
Means	15.02	65.65	5.37	14.50	21.63	7.13
S.D.	0.951	24.27	2.59	7.53	10.23	6.18

TABLE XIII

INTERCORRELATIONS OF VARIABLES (COURSE IV)

	Age	P.R.	S.A.	P.T.	F.	G.
Age	......					
P.R.	-.175	......				
S.A.	0.044	-.521	......			
P.T.	-.122	0.609	-.605	......		
F.	-.011	0.557	-.667	0.799	......	
G.	0.130	0.180	-.366	0.105	0.672	......

are coded by number. High score indicates attainment of desired objectives. Symbol for initial test score is P. T.

7. Final scores on certain standardized tests. Same as above with F to signify final. Scored as number 6.

8. Pupil Age measured in completed integral units. Symbol A.

9. Pupil rank in intelligence. Symbols used are P. R. for percentile ranks on statewide testing program involving Henmon-Nelson Test of Mental Ability and I. Q. for intelligence quotient on Otis form A.

10. Average scholastic standing in high school. Symbol is S. A.; scored on scale of 1 to 12, with 1 corresponding to A, 2 to A-, 3 to B+, ..., 12 to F.

From items 6 through 10 based upon test, grades, and school records it was hoped that three criteria of teaching success could be developed. The test criteria are based on discrepancies between actual and expected scores on the pre-test, final and gains. These three were to be compared separately with the rating of items 1 through 5 by usual correlational techniques. Since the development of criteria of teaching effectiveness based on pupil achievement involves considerable detail and computation, priority is given to that treatment.

The means and standard deviations of the ages, percentile ranks in intelligence or intelligence quotients, average scholastic standings, initial and final test scores on the achievement examinations, and gains or adjusted gains are listed in Tables III, VI, IX, XII, XV, XVIII, and XXI. The number of cases on which each of these statistics was computed is given in each table.

The intercorrelations between these variables are given for each course in Tables IV, VII, X, XIII, XVI, XIX, and XXII. For some of the variables in Table XXII, two sets of correlations are given. One of the participating classes was tested only one time. The upper correlation in each cell is based on 103 cases while the lower used only 83. The correlations for the 103 were used in preparing the prediction equations for the initial scores (P. T. ') for course VII. The final and gain predictions for that course are based on the 83 cases.

Beta coefficients for prediction equations in standard score form were computed from the intercorrelations of each course group by the Aitkin Method of Pivotal Condensations (62:89-95). The standard score equations were subsequently converted to the raw score form of the regression equation by replacing the means and standard deviations in the definition of standard score by the computed values given in the tables referred to above. Prediction equations in raw score form are listed in Tables V, VIII, XI, XIV, XVII, XX, and XXIII.

At the time that the beta coefficients were computed, multiple correlations were also calculated. If one adds variables to an equation of this type, one expects the multiple correlation to increase. If it does not, the added variable has not improved the prediction that can be made by the use of the regression equation. In this computation, if the multiple correlation was not raised by an increment of 0. 02 on the addition of a variable, the variable was omitted from the equation. Consequently, some equations have only two independent variables while others have three or four. If this general procedure failed to raise the multiple correlations above 0. 3, the regression equation was not set up in raw score form because it fails to take account of sufficient variance to make the operation worth the effort involved.

As a result of the consideration mentioned in the preceding paragraph, only three prediction equations were set up for predicting gains. Later it was decided to use only such equations whose multiple correlations exceeded 0. 5. This eliminated all but one gain predictor. Since this one did not differ materially from 0. 5 it was decided to omit it also. All the multiple correlations for the various prediction equations are given in Table XXIV.

Although it was not feasible to measure teaching effectiveness on the discrepancy between the actual gain and the predicted gain score, it was felt that it would be justifiable to base judgment on the gains alone, inasmuch as gains are the outcomes that are desired. Evidently none of the measures used correlated highly enough with gain to possess predictive value. Since we do not have the necessary means for predicting gains, it is assumed that gains are equally likely to appear in each class if we disregard the influence of the teacher. Accordingly the assessment of teaching effectiveness in terms of actual gains or gains adjusted for pretest handicap was substituted for the proposed criterion of the discrepancy between actual and predicted mean gains.

The mean discrepancies for the different classes between the pretest and predicted pretest score are summarized in Table XXV. The same type of information concerning the final score discrepancies is located in Table XXVI. The letter D is used to indicate the discrepancy while the subscripts p and f refer to the pretest and final respectively.

When the means of the gains for each each pretest score were graphed versus the pretest scores, it became apparent that definite relationships existed between pretest and gain for Courses II, III, VI, and VII. In all of these courses high gains tended to be associated with low pretest scores. Accordingly, the baseline

TABLE XIV

PREDICTION EQUATIONS FOR COURSE IV

P. T.'	=	0. 125 P. R.	-1. 15 S. A.	+12. 47	
F.'	=	0. 018 P. R.	-1. 10 S. A.	+ 0. 82 P. T.	+14. 44
G.'	=	- . 004 P. R.	- . 89 S. A.	+12. 14	

TABLE XV

MEANS AND STANDARD DEVIATIONS OF VARIABLES (COURSE V)
(N = 105)

	Age	P. R.	S. A.	P. T.	F.	G.
Means	14. 39	55. 92	5. 43	22. 49	29. 09	6. 60
S. D.	0. 67	26. 02	2. 45	9. 14	12. 53	7. 95

TABLE XVI

INTERCORRELATIONS OF VARIABLES (COURSE V)

	Age	P. R.	S. A.	P. T.	F.	G.
Age	......					
P. R.	- . 297	......				
S. A.	0. 356	- . 564	......			
P. T.	- . 093	0. 424	- . 481	......		
F.	- . 144	0. 521	- . 514	0. 774	......	
G.	- . 119	0. 333	- . 256	0. 070	. 686	......

TABLE XVII

PREDICTION EQUATIONS FOR COURSE V

P. T.'	=	0. 079 P. R.	-1. 317 S. A.	+25. 24	
F.'	=	0. 093 P. R.	- . 475 S. A.	+ . 887 P. T.	+6. 51
G.'	=	0. 085 P. R.	- . 324 S. A.	+ 3. 63	

TABLE XVIII

MEANS AND STANDARD DEVIATIONS OF VARIABLES (COURSE VI)
(N = 147)

	Age	P. R.	S. A.	P. T.	F.	G.
Means	13.95	49.84	5.85	87.48	94.61	7.04
S. D.	0.54	25.85	2.48	15.00	12.98	12.24

TABLE XIX

INTERCORRELATIONS OF VARIABLES (COURSE VI)

	Age	P. R.	S. A.	P. T.	F.	G.
Age	......					
P. R.	- .196	......				
S. A.	0.178	- .475	......			
P. T.	- .152	0.534	- .562	......		
F.	- .208	0.549	- .565	0.790	......	
A. G.	- .112	0.146	- .103	- .041	0.442	......

TABLE XX

PREDICTION EQUATIONS FOR COURSE VI

P. T.'	= 0.200 P. R.	-2.41 S. A.	+91.58	
F.'	= 0.072 P. R.	- .73 S. A.	+ .550 P. T.	+47.24
A. G.'	not used; cannot account for 10 percent of variance			

TABLE XXI

MEANS AND STANDARD DEVIATIONS OF VARIABLES (COURSE VII)

	Age	P. R.	S. A.	P. T.	F.	A. G.
Mean	16.37	50.00	5.69	93.50	101.02	7.34
S. D.	0.74	28.85	2.29	18.43	16.78	14.71
N	83	83	83	83	83	83
Mean	16.36	48.86	5.71	94.71		
S. D.	0.74	29.34	2.30	18.72		
N	103	103	103	103		

TABLE XXII

INTERCORRELATIONS OF VARIABLES (COURSE VII)

	Age*	P. R. *	S. A. *	P. T.	F.	A. G.
Age	......					
P. R. *	- .311 - .340	......				
S. A. *	0.286 0.326	- .541 - .520	......			
P. T. *	- .281 - .280	0.490 0.484	- .542 - .518	......		
F.	- .408	0.607	- .595	0.858	......	
A. G.	- .060	0.191	- .098	- .063	0.271	......

*Two sets of correlations given; upper number based on 103 cases, lower set based on 83 cases.

TABLE XXIII

PREDICTION EQUATIONS FOR COURSE VII

P. T. ' = . 177 P. R. -3. 190 S. A. +104. 260

F. ' = 2. 743 Age +0. 105 P. R. - . 799 S. A. +0. 620 P. T. +87. 260

A. G. not used; cannot account for ten percent of variance.

TABLE XXIV

CALCULATED MULTIPLE CORRELATIONS FOR REGRESSION EQUATIONS

Course	P. T.	F.	G.
I	0. 64	0. 87	0. 53
II	0. 51	0. 66	*
III	0. 80	0. 87	*
IV	0. 70	0. 83	0. 35
V	0. 51	0. 80	0. 34
VI	0. 64	0. 81	*
VII	0. 59	0. 90	*

*Indicates multiple correlations less than 0. 3

TABLE XXV

MEAN DISCREPANCIES BETWEEN PRETEST FOR EACH CLASS AND PREDICTED PRETEST

Course	Teacher	D_p	Course	Teacher	D_p
I	27	- .81		19	+2.52
	A	+ .93		16	- .50
	B	+1.34	V	14	- .11
	28	- .66		16	- .87
	29a	- .94		12	- .73
	29b	-1.26		15	+3.25
II	20	+1.00		13+G	-3.49
	22	-1.23	VI	4	+7.36
	23	-1.50		7	-1.90
	24	- .98		3	-1.73
	25	+1.07		1	-1.97
	26	+ .85		8	-2.73
III	31	+3.97		5	+ .22
	C	- .23	VII	11	+6.12
	32	-3.74		1	+7.70
	D	-1.55		21	+2.27
	E	+ .08		10	-1.63
IV	18	+1.10		9	-5.81
	13+F	-4.44		6	+ .28
	17	- .27			

TABLE XXVI

MEAN DISCREPANCY BETWEEN FINAL AND PREDICTED FINAL FOR EACH CLASS

Course	Teacher	D_f	Course	Teacher	D_f
I	27	+ .89		17	+0.73
	A	-1.27		19	+1.91
	B	+1.91		16	+0.67
	28	+2.33	V	14	-6.00
	29a	-2.61		16	-0.71
	29b	-1.16		12	+1.91
II	20	-0.38		15	+3.75
	22	+0.73		G	-2.20
	23	+0.23	VI	4	+2.65
	24	-0.98		7	+2.32
	25	+0.77		3	-2.54
	26	-0.12		1	+4.54
III	31	-0.86		8	-1.67
	C	-0.68	VII	5	-3.59
	32	+0.20		1	+0.74
	D	+0.38		21	+1.03
	E	+1.12		10	+1.44
IV	18	-1.92		9	-2.10
	F	-1.31		6	+0.28

TABLE XXVII

MEAN GAINS OR ADJUSTED MEAN GAINS FOR EACH CLASS

Course	Teacher	G.	Course	Teacher	G.
I	27	6.40		17	7.40
	A	3.00		19	9.52
	B	5.87		16	8.55
	28	8.75	V	14	0.28
	29a	0.00		16	5.52
	29b	2.53		12	8.59
II	20	6.40		15	11.21
	22	6.62		G	5.90
	23	7.52	VI	4	10.73
	24	5.98		7	9.37
	25	5.89		3	5.97
	26	4.77		1	12.33
III	31	8,49		8	4.50
	C	6.30		5	1.79
	32	9.70	VII	1	4.87
	D	8.39		21	10.77
	E	9.92		10	6.72
IV	18	4.34		9	6.07
	F	5.81		6	9.89

TABLE XXVIII

MEANS AND STANDARD DEVIATIONS OF MEAN DISCREPANCIES AND GAINS FOR EACH COURSE

	Discrepancies on Pretest		Discrepancies on Final		Mean Class Gains	
Course	Mean	S. D.	Mean	S. D.	Mean	S. D.
I	- .23	0.99	- .02	1.81	4.43	2.88
II	0.13	1.12	0.04	0.62	6.20	0.83
III	- .29	2.52	0.03	0.73	8.56	1.29
IV	- .31	2.33	0.02	1.41	7.12	1.86
V	- .39	2.15	- .65	3.38	6.30	3.64
VI	- .14	3.46	0.29	3.03	7.45	3.68
VII	0.75	5.03	0.28	1.25	7.66	2.26

of the pretest scores (abscissa) was divided into segments. A mean gain ordinate was calculated for each segment. Each gain score in the course was multiplied by the ratio of the mean gain for the segment of the abscissa which contains its pretest score. Thus each gain score in these four courses was adjusted for inequalities in difficulty at various levels on the test. Table XXVII shows the mean gains for classes in Courses I, IV, and V and the adjusted mean gains for each class in Courses II, III, VI, and VII.

The discrepancy and mean gain scores serve to rank the teachers on these particular criteria in the various courses. A common device for comparing groups of this kind is to reduce the scores to standard score form with zero mean and unit standard deviation. Any score in a group of scores can be converted to standard score form by subtracting the mean of the group and dividing this difference by the standard deviation of the group. The means and standard deviations of the discrepancy and mean gains of all classes for all courses are listed in Table XXVIII. When these means and standard deviations were used to convert the discrepancy and gain scores into standard scores, the standard scores tabulated in Table XXIX were obtained for the three criteria of teaching effectiveness.

Standard score distributions always contain negative as well as positive numbers. Under ordinary circumstances the signed numbers are perfectly acceptable. However, in this investigation a weighting correction was made to compensate for the fact that possibly the more capable teachers may be teaching in one course while mediocre teachers are clustered in another. The plan for weighting was to multiply the score for each teacher by the ratio of the mean score for all pupils in that course to the mean score as reported for the norms. Such a plan would project the lowest score in a group down in proportion to the rating of the group on the norms. However, if a constant is added which will change all the standard scores to positive values, this objection is removed. A constant of five was added to each standard score. The resulting scores are listed in Table XXX.

Table XXXI shows the means attained by the pupils taught by the teachers being evaluated on the standardized achievement examinations. The means of the norms are furnished for comparison. When the scores of Table XXX were multiplied by the ratio of the appropriate means from Table XXXI, the scores were changed to criteria scores as listed in Table XXXII.

The intercorrelations of these three criteria scores and the five ratings by principal, teacher, pupils, peers, and administrative criterion X are summarized in Table XXXIII. These correlations were computed by the Pearson product moment formula. Since low numbers on the rating scales indicated superior teaching and high numbers superior teaching on the achievement criteria, the signs of correlations obtained using one criterion from each group have been reversed to give the usual interpretation to the correlations.

If samples of size thirty had been randomly drawn from an infinite population whose true correlation is zero, the standard error would be 0.18. Only in about five percent of such samples would one expect to find correlations whose absolute value is greater than 0.36. In this set of correlations based on a selected group, only seven correlations are of the magnitude to cast doubt upon the possibility of their having arisen as chance fluctuations from a true correlation of zero. The seven are:

1. Correlation of M-Blank with Administrative Criterion X
2. Self rating with principals M blank
3. Peer with pupil
4. Peer with Administrative Criterion X
5. Pretest rating with final rating on pupil achievement
6. Pretest with adjusted gain criterion
7. Final with adjusted gain

There may be many who question the assumption that the means on the norms furnished by standardized test publishers are equally applicable to all these teachers. It may be argued that the intercorrelations among these criteria based on pupil achievement are spuriously high in that the ratios of the means move subgroups of the entire group in the same direction for all three criteria. In consideration of this valid objection the writer calculated the correlations among the criteria scores listed in Table XXX. This uses the assumption that the teachers in every class are comparable with those of all the other classes. Using this procedure the correlations obtained are as follows:

1. Pretest with final 0.16
2. Pretest with adjusted gain -.01
3. Final with gain 0.69

The only one of these that could be considered significantly different from zero is the correlation of final score with gain.

As a matter of interest it was decided to check the consistency of the ratings by pupils and peers. In the data sheet the ratings by the pupils are numbered from one to five. These are numbers that were assigned to the rating sheets before they were sent to the pupils. The ratings in column Pu_1 represent the ratings of their respective teachers by the pupils who re-

TABLE XXIX

STANDARD SCORES FOR TEACHERS IN COURSES ON THE THREE CRITERIA

Course	Teacher	Z_p	Z_f	Z_g
I	27	-0.59	+0.49	+0.69
	A	+1.17	-0.71	-0.49
	B	+1.60	+1.05	+0.51
	28	-0.43	+1.28	+1.52
	29a	-0.72	-1.45	-1.56
	29b	-1.04	-0.65	-0.66
II	20	+1.01	-0.68	+0.24
	22	-0.98	+1.11	+0.51
	23	-1.22	+0.31	+1.59
	24	-0.76	-1.65	-0.27
	25	+1.07	+1.18	-0.37
	26	+0.88	-0.26	-1.72
III	31	+1.69	-1.22	-0.04
	C	+0.02	-0.97	-1.75
	32	-1.32	+0.23	+0.88
	D	-0.50	+0.48	-0.13
	E	+0.15	+1.49	+1.05
IV	18	+0.61	-1.38	-1.49
	13 and F	-1.77	-0.94	-0.70
	17	+0.01	+0.50	+0.15
	19	+1.21	+1.34	+1.29
	16	-0.08	+0.46	+0.77
V	14	+0.13	-1.58	-1.65
	16	-0.22	-0.02	-0.21
	12	-0.15	+0.76	+0.63
	15	+1.69	+1.30	+1.35
	13 and G	-1.44	-0.46	-0.11
VI	4	+2.17	+0.78	+0.89
	7	-0.50	+0.67	+0.52
	3	-0.46	-0.93	-0.40
	1	-0.53	+1.40	+1.33
	8	-0.75	-0.64	-0.80
	5	+0.10	-1.28	-1.54
VII	11	+1.07	.....	.....
	1	+1.38	+0.37	-1.23
	21	+0.30	+0.60	+1.38
	10	-0.47	+0.93	-0.41
	9	-1.30	-1.90	-0.70
	6	-0.97	0.00	+0.99

TABLE XXX

STANDARD SCORES OF TEACHERS WITH CONSTANT FIVE ADDED

Course	Teacher	S_p	S_f	S_g
I	27	4.41	5.49	5.69
	28	4.57	6.28	6.52
	29*	4.12	3.95	3.89
II	20	6.01	4.32	5.24
	22	4.02	6.11	5.51
	23	3.78	5.31	6.59
	24	4.24	3.35	4.73
	25	6.07	6.18	4.63
	26	5.88	4.74	3.28
III	31	6.69	3.78	4.96
	32	3.63	5.23	5.88
IV	18	5.61	3.62	3.51
	13**	3.23	....	....
	17	5.01	5.50	5.15
	19	6.21	6.34	6.29
	16	4.92	5.46	5.77
V	14	5.13	3.42	3.35
	16	4.78	4.98	4.79
	12	4.85	5.76	5.63
	15	6.69	6.30	6.35
	13**	3.56	....	....
VI	4	7.17	5.78	5.89
	7	4.50	5.67	5.52
	3	4.54	4.07	4.60
	1	4.47	6.40	6.33
	8	4.25	4.36	4.20
	5	5.10	3.72	3.46
VII	11**	6.07	....	....
	1	6.38	5.37	3.77
	21	5.30	5.60	6.38
	10	4.53	5.93	4.59
	9	3.70	3.10	4.30
	6	4.03	5.00	5.99

---

* Average of scores for two sections

**Can be evaluated over one semester only. Only pretest used.

TABLE XXXI

COMPARISON OF MEANS OF NORMING AND COURSE POPULATIONS

Course	Course Means	Norm Means
I	17.34	21.9
II	9.23	10.0
III	54.54	47.65
IV	21.63	21.9
V	29.17	20.2
VI	91.04	89.0
VII	97.87	104.0

TABLE XXXII

TEACHER SCORES ON ACHIEVEMENT CRITERIA AFTER WEIGHTING

Course	Teacher	C_p	C_f	C_g
I	27	3.49	4.35	4.51
	28	3.62	4.97	5.16
	29	3.26	3.13	3.08
II	20	5.55	3.99	4.84
	22	3.71	5.64	5.09
	23	3.49	4.90	6.08
	24	3.91	3.09	4.37
	25	5.60	5.70	4.27
	26	5.43	4.38	3.03
III	31	7.66	4.33	5.68
	32	4.15	5.99	6.73
IV	18	5.54	3.58	3.47
	13	3.19	....	....
	17	4.95	5.43	5.09
	19	6.13	6.26	6.21
	16	4.86	5.39	5.70
V	14	7.39	4.93	4.83
	16	6.89	7.17	6.90
	12	6.99	8.30	8.11
	15	9.64	9.08	9.15
	13	5.13	....	....
VI	4	7.33	5.91	6.02
	7	4.60	5.80	5.64
	3	4.64	4.16	4.70
	1	4.57	6.54	6.47
	8	4.35	4.46	4.29
	5	5.21	3.80	3.54
VII	11	5.71	....	....
	1	6.00	5.05	3.55
	21	4.99	5.27	6.00
	10	4.26	5.58	4.32
	9	3.48	2.92	4.05
	6	3.79	4.71	5.64

TABLE XXXIII

INTERCORRELATIONS OF CRITERION SCORES OF TEACHERS

	M	Pu	Se	Pe	AX	PT	F.	AG.
M	.....							
Pu	0.13	.....						
Se	0.43	0.30	.....					
Pe	0.29	0.38	0.23	.....				
AX	0.57	0.22	0.33	0.40	.....			
PT.	- .04	0.02	- .17	- .02	0.09	.....		
F.	0.02	0.13	- .18	0.23	0.00	0.59	.....	
AG.	0.10	0.11	- .15	0.23	0.10	0.54	0.85	.....

TABLE XXXIV

RATINGS OF THE ACHIEVEMENT TESTS BY THE TEACHERS

Test Area	Rating by Teacher	Criticisms or Opinions About Tests
Mathematics	2 2	None given in this area.
World History	3 3 4 2	1. To comprehensive—covers materials not studied. 2. Doesn't measure the ability to reason from facts of history.
Typing	2 2 4 5 2	1. Speed and accuracy are not only aims in typing. 2. Tension results when the principal administers test.
Algebra	2 * 2 2	1. Goals other than subject matter should be emphasized. * letter without rating
Science	2 1 or 3* 2	1. Too much material not covered in class * 1 on materials covered; 3 on other items
English	2 3 3 1 1 2 3 2 2	1. Does not test ability to read for basic ideas and interpretation. 2. Lack of opportunity for individual expression. 3. Absence of evaluation of literature. 4. No test for written expression. 5. Too much emphasis on grammar.

ceived the blanks numbered one. The other columns have ratings which were obtained in a similar manner. The scores of the teachers on the five pupil ratings were correlated. This produced ten correlations, each of them based on the judgments made by 52 pupils about 26 teachers. The seperation of the pupils into groups was predetermined as the blanks were sent out. The range of the ten correlations was 0.00 to 0.74 with a median value of 0.42. In a comparable manner the ratings of the teachers by two groups of teaching colleagues were correlated (r = 0.26). These two groups were chosen as far as possible to represent the subject matter area in which the teacher was working. An attempt was made to divide the two groups into men and women when the two original forms were sent out. The consistencies in rating, as measured here, are quite low.

Any evaluation of teaching effectiveness by pupil achievement on standardized tests is invalid insofar as the tests fail to measure the objectives of instruction. In order to check on this aspect of the validity of the testing program, the reaction of each teacher to the tests used in his classes were sought by means of a short questionnaire. (Appendix II. H) The teachers were asked to rate the test on a five point scale with 1 indicating a very satisfactory test and 5 very unsatisfactory. The teachers were also invited to make any criticisms or comments they felt appropriate. Of the thirty-two teachers invited to rate and comment, twenty-seven responded. Only a few did more than rate the test and offer some brief comment. No second attempt was made to secure the five additional ratings because the school year was almost over. The judgments and comments of the teachers are tabulated by classes in Table XXXIV. Only those comments which may be construed as criticisms are tabulated because the original questionnaire sought the criticisms only. In general, the criticisms are confined to two categories: (1) tests cover materials not taught in classes; and (2) important objectives are excluded. The mean value of the ratings of the tests (2.4) is indicative that the tests were slightly better than adequate.

When the intercorrelations among the three measures of teaching effectiveness based on pupil achievement are considered after weighting for differences in accomplishment in the various courses, there definitely appears to be a positive correlation among them. The standard error for samples of size thirty is 0.18 when randomly drawn from an infinite population. All the intercorrelations among the pretest, final, and gain criteria are at least three times as large. It would appear unreasonable to believe that the true correlations are zero. As pointed out previously, the correlations drop so low that they might be regarded as chance fluctuations from zero for all except the final with gain criterion (r = 0.69) if the justification of weighting groups according to their means is denied.

When the principal's evaluation on the M-Blank was correlated with the other measures of teaching effectiveness, it was found to correlate 0.57 with the administrative X criterion and 0.43 with the teacher's self rating. All the other correlations may be readily explained as possible chance variations from a true zero. The sum of the pupil ratings was found to correlate 0.38 with the sum of the evaluations of the peers. No other significant correlations were found for this variable. Peer rating correlated 0.40 with the administrative X criterion. Self-evaluation was found to correlate 0.33 with the administrative X criterion. The other relationships are too small to warrant mention.

In summary, it would appear from the results of this investigation that no appreciable relationships exist between rating criteria and pupil attainment criteria. Evidently the criteria evaluate different aspects of teaching effectiveness. There are low positive correlations among the constituent parts of each of these two groups of criteria. Whether one uses a composite criterion or several separate criteria will depend upon the purpose of the evaluation undertaken. It is not easy, even if sometimes desirable, to represent effectiveness in a complex activity such as teaching, by a single criterion.

SECTION V

SUMMARY, CONCLUSIONS. AND IMPLICATIONS

Summary and Conclusions

THE PURPOSE of this investigation was to determine the relationships among certain criteria used to evaluate the teaching effectiveness of graduates of the School of Education, University of Wisconsin. The teaching effectiveness of a selected group was estimated by the use of eight different criteria. The relationships among the specific criteria were expressed by linear coefficients of correlation.

The group studied in this investigation consisted of thirty teachers who received the University Teachers Certificate in 1950 and are now employed in twenty-eight high schools in the State of Wisconsin. This selected group appears to differ from the universe of graduates of the School of Education in that they (1) exhibit a slight superiority in grade point average, (2) have one year of teaching experience, and

(3) are teaching in selected subject matter areas. With these exceptions, there is no reason for not believing the group representative of the graduates of the School of Education.

The subjects were evaluated on separate types of rating scales by principals, peers, pupils, the teacher himself, and an outside agency whose identity is confidential. In addition, the subject matter achievements of their pupils, as measured on certain standardized tests, were used to develop three other criteria of teaching effectiveness.

Complete data, including measures of intelligence, age, and average scholastic standing, as well as the scores on the initial and final administration of the achievement examinations, were secured for seven hundred and forty-one pupils taught by these teachers. On the basis of these data, the teachers were evaluated as to initial test status, final test status, and pupil gain. Pretest status was justified as a measure of teaching effectiveness in that three months of the school year had elapsed prior to the testing.

It was assumed that the achievement of the pupils is a function of (1) the effects of the teacher, (2) the intelligence, age, average scholastic standing, and previous knowledge of the subject, and (3) other factors not mentioned. Regression equations were set up which predicted the achievement to be expected on the basis of the measured pupil factors. When these predicted achievements were subtracted from the actual pupil achievements, the residuals were considered to be the results of the efforts of the teachers plus the effects of the unmeasured factors. It would seem reasonable to assume that the contributions of the unmeasured factors were approximately equally distributed among the classes taught by these teachers. Such residual estimates were obtained for each class using both pretest and final class standings. Since none of the measured pupil characteristics could be used for predicting pupil gain, the gains themselves were used as measures of effectiveness. In some courses the gains were, however, adjusted for difficulties associated with high pretest scores.

The standings of the thirty teachers on the eight criteria were correlated by the Pearson product moment method. No correlation appreciably different from zero was discovered between the evaluations of the teachers on the different rating scales and the evaluations based on the achievements of their pupils in the subject matter areas. Although low, there appears to be a positive association among the assessments of teaching ability made by the principals, pupils, peers, teachers, and the administrative X criterion. A considerable amount of relationship seems to exist between the evaluations of the teachers based on the final achievement of their pupils in tests of subject matter when differences in age, intelligence, average scholastic standing, and pretest knowledge are partialed out and the evaluations based on pupil gain in subject matter after the adjustment for differences in initial score.

Discussion and Implications

The results of this study imply that the evaluation of teachers is to some extent influenced by the choice of the criterion. Evaluation must be made in terms of the criterion used as well as the rater, the situation, and other factors.

If a thorough knowledge of the effectiveness of the teacher is desired, the several criteria must be used. If all of these criteria are thought to have value, none correlated highly enough with the other to justify omission on the grounds that they repeat the effects of the other criteria. Presumably the validation of all criteria must, in the final analysis, rest with the philosophy adhered to.

In some instances a composite evaluation may seem desirable. To employ simple averages assumes that all components are equally weighted, an assumption that probably is very questionable. Accordingly, some future research must use factor analysis or other statistical technique to properly weight the components.

The present study assumed that pupil gain in subject matter was an important and practical approach to criteria building. If such an approach is used, it would seem advisable that future studies be designed to investigate the relationship between subject matter gain and gain in the more intangible aspects of the desired concomitant learnings. Certainly there is a need of developing more reliable and valid instruments for measurement of the concomitant learnings.

If pupil gains are to be used as a measure of teaching effectiveness, the tests must be given under such conditions that the gains are much larger than the standard error of the tests. This would require that the tests measure the objectives of instruction and that the period between test and retest be sufficiently long to permit large gains. A casual examination of the gains produced and the standard errors as given in the norms serves to advocate caution in the interpretation of gain scores as a criterion. While the tests employed in this study have high reliability one may question whether they covered the objectives of the individual teacher.

Rating scales when used as criteria lack objectivity. If, for example, the same rating scale is used by two individuals, the correlations, although positive, are usually low. This was amply demonstrated by the correlations of the five groups of pupil ratings and the peer ratings as used in this investigation. If teacher ratings are to be used, this objectivity and the reliability need to be improved.

Possibly raters tend to overrate instructors when using rating scales. A study of the data for the rating scales employed shows that the mean rating places Wisconsin graduates considerably above the mean for the teachers in participating schools. While those who train teachers at Wisconsin would prefer such an interpretation, it is more probable that it represents the ego of the raters and also the amount of personal relationship between the rater and the ratee.

In general, no adequate basis for validation of teacher evaluations exist at present. There is apparently no general agreement as to what is good teaching, and even if there were, present day measures lack the reliability necessary for valid criteria.

BIBLIOGRAPHY

1. Adams, W. C. T. "Teacher Rating Card," American School Board Journal, LVIII (December 1919), p. 30.
2. Albert, H. R. "An Analysis of Teacher Rating by Pupils in San Antonio, Texas," Educational Administration and Supervision, XXVII (April 1941), pp. 267-74.
3. Baier, Donald E. "Reply to Travers: A Critical Review of the Validity and Rationale of the Forced Choice Technique," Psychological Bulletin, XLVIII (September 1951), pp. 421-34.
4. Barr, A. S. and others. "The Validity of Certain Instruments Employed in the Measurement of Teaching Ability," in The Measurement of Teaching Efficiency, Helen M. Walker, Editor (New York: Macmillan Co., 1935). pp. 71-141.
5. ——— "Measurement of Teaching Ability," Review of Educational Research (June 1934), pp. 261-67; 329-30.
6. ——— "Measurement of Teaching Ability," Review of Educational Research X (June 1940), pp. 182-84.
7. ——— "The Measurement and Prediction of Teaching Efficiency," Review of Educational Research, XIII (June 1943), pp. 218-23.
8. ——— "The Measurement of Teacher Characteristics and the Prediction of Teaching Efficiency," Review of Educational Research, XXII (June 1952), pp. 169-73.
9. Beecher, Dwight E. Evaluation of Teaching (Syracuse, N. Y.: Syracuse University Press, 1949), 105 pp.
10. Betts, Gilbert L. Reliability and Validity of Measures of Teaching Ability or Teaching Success, Special Survey Studies, National Survey of the Education of Teachers, Bulletin, 1933, No. 10, Vol. V, Part II (Washington, D. C.: Government Printing Office, 1935), pp. 87-153.
11. Bimson, Oliver H. "Do Good Teachers Produce Good Results?" North Central Association Quarterly, XII (October 1937), pp. 271-76.
12. Boardman, Charles W. Professional Tests as Measures of Teaching Efficiency in High School, Contributions to Education, No. 327 (New York: Teachers College, Columbia University, 1928), 85 pp.
13. Book, M. F. "The High School Teacher from the Pupils' Point of View," Pedagogic Seminary, XII (September 1905), pp. 239-88.
14. Boyce, A. C. "Methods of Measuring Teachers Efficiency," Part II, Fourteenth Yearbook of N. S. S. E. (Bloomington, Ill.: Public School Publishing Co., 1915).
15. Boyce, Robert B., and Bryan, Roy C. "To What Extent Do Pupils' Opinions of Teachers Change in Later Years?" Journal of Educational Research, XXXVII (May 1944), pp. 698-705.
16. Brandt, W. J. "Followup of Some Earlier Wisconsin Studies of Teaching Efficiency," Journal of Experimental Education, XVIII (September 1949), pp. 1-29.
17. Brooks, Samuel S. "Measuring the Efficiency of Teachers by Standardized Tests," Journal of Educational Research, IV (November 1921), pp. 255-64.
18. Bryan, Roy C. "Pupil Rating of Secondary School Teachers," School Review, XLVI (May 1938), pp. 357-67.
19. Butsch, R. L. C. "Teacher Rating," Review of Educational Research, I (April 1931), pp. 99-107.
20. Campbell, R. F. "Evaluation and the Rating of Teachers," Elementary School Journal, XLI (May 1941), pp. 671-76.
21. Cook, Walter W., and Leeds, Carroll H. "Measuring the Teaching Personality," Educational and Psychological Measurement VII (May 1947), pp. 399-410.
22. Cook, William H. "Uniform Standards for Judging Teachers in South Dakota," Educational Administration and Supervision, VII (1921), pp. 1-11.
23. Coy, Genevieve L. "A Study of Various Factors Which Influence the Use of the Accomplishment Quotient as a Measure of Teaching Efficiency," Journal of Educational Research, XXI (January 1930), pp. 29-42.
24. Crabs, Lelah M. Measuring Efficiency in Supervision and Teaching, Contributions to Education, No. 175 (New York: Teach-

ers College, Columbia University, 1925), 98 pp.

25. Cronbach, Lee J. "Test Reliability: Its Meaning and Determination," Psychometrika, XII (March 1947), pp. 1-17.
26. Domas, S. J., and Tiedeman, D. V. "Teacher Competence: An Annotated Bibliography," Journal of Experimental Education, XIX (December 1950), pp. 101-218.
27. Evans, Kathleen M. "A Critical Survey of Methods of Assessing Teaching Ability," British Journal of Educational Psychology, XXI, Part II, (June 1951), pp. 89-95.
28. Goetsch, E. W. "In Service Checkup of Iowa State Teachers College Graduates," School and Society, LXIX (June 1949), pp. 419-23.
29. Good, Carter V., and others. The Methodology of Educational Research (New York: D. Appleton-Century Co., 1941), 863 pp.
30. Gotham, R. E. "Personality and Teaching Efficiency," Journal of Experimental Education, XIV (December 1945), pp. 157-65.
31. Guthrie, E. R. "Evaluation of Faculty Service," American Association of University Professors Bulletin, XXXI (June 1945), pp. 255-62.
32. Hampton, N. D. "An Analysis of Supervisory Ratings of Elementary Teachers Graduated from Iowa State Teachers College," Journal of Experimental Education, XX (December 1951), pp. 179-215.
33. Hansen, Basil C. "Open the Curtain on this Farce Called Teacher Evaluation," Educational Administration and Supervision, XXXII (October 1946), pp. 412-18.
34. Hielman, J. D., and Arman Trout, W. D. "The Rating of College Teachers on Ten Traits by Their Students," Journal of Educational Psychology, XXVII (March 1936), pp. 197-216.
35. Hildreth, Gertrude. "Results of Repeated Measurement of Pupil Achievement," Journal of Educational Psychology, XXI (April 1930), pp. 286-96.
36. Hill, C. W. "The Efficiency Ratings of Teachers," Elementary School Journal, XXI (February 1921), pp. 438-43.
37. Hoppoch, Robert. "New York University Students Grade Their Professors," School and Society, LXVI (July 1947), pp. 70-72.
38. Johnson, Palmer O. Statistical Methods in Research (New York: Prentice Hall, 1949), 357 pp.
39. Jones, Ronald D. "The Prediction of Teaching Efficiency from Objective Measures," Journal of Experimental Education, XV (September 1946), pp. 85-99.
40. Knight, F. G. Qualities Related to Success in Teaching (New York: Teachers College, Columbia University, 1922), 67 pp.
41. Kratz, Henry Elton. Studies and Observations in the Schoolroom (New York: Educational Publishing Co., 1907), 220 pp.
42. LaDuke, C. V. "The Measurement of Teaching Ability, Study No. 3," Journal of Experimental Education, XIV (September 1945), pp. 75-100.
43. Lindquist, E. F., and others. Educational Measurement (Washington, D. C.: American Council on Education, 1951), 808 pp.
44. Lins, Leo J. "The Prediction of Teaching Efficiency," Journal of Experimental Education, XV (September 1946), pp. 2-60.
45. McCartha, Carl W. "The Practice of Teacher Evaluation in the Southwest in 1948," Journal of Educational Research, XLIV (October 1950), pp. 122-28.
46. McNemar, Quinn. Psychological Statistics (New York: John Wiley and Sons, Inc., 1949), 344 pp.
47. Merriam, J. L. Normal School Education and Efficiency in Teaching (New York: Teachers College, Columbia University, 1906).
48. Myers, G. C. "Teachers' Ratings," School and Society, V (March 17, 1917), pp. 322-23.
49. Peters, C. C., and VanVoorhis, W. R. Statistical Procedures and Their Mathematical Bases (New York: McGraw Hill Book Co., 1940), 478 pp.
50. Pittenger, B. F. "Problems in Teacher Measurement," Journal of Educational Psychology, VIII (1917), pp. 103-10.
51. Remmers, H. H., and others. "Interrelationship of Various Teaching Criteria," American Psychologist (abstract), IV-7 (July 1949), p. 288.
52. Rider, Paul R. An Introduction to Modern Statistical Methods (New York: John Wiley and Sons, Inc., 1939), 192 pp.
53. Rolfe, J. G. "The Measurment of Teaching Ability, Study No. 2," Journal of Experimental Education, XIV (September 1945), pp. 52-74.
54. Ross, C. C. Measurement in Today's Schools (New York: Prentice-Hall, 1947), 532 pp.
55. Rostker, Leon E. "The Measurement of Teaching Ability, Study No. 1," Journal of Experimental Education, XIV (September 1945), pp. 6-51.
56. Ruediger, W. C., and Strauer, G. D. "The Qualities of Merit in Teachers," Journal of Educational Psychology, I (1910), pp. 272-78.
57. Rugg, Harold. "Is the Rating of Human Character Possible?" Journal of Educational Psychology, XII (November, December 1921), pp. 425-38; 485-501; XIII

(January, October 1922), pp. 30-42; 81-93.

58. Ryans, David G. "The Criteria of Teaching Effectiveness," Journal of Educational Research, XLII (May 1949), pp. 690-99.

59. Schultz, Frank G. "Appraising and Rewarding Teaching Effectiveness," Charting the Course for American Higher Education in a Period of Partial Mobilization (Washington, D. C.: Department of Higher Education, National Education Association, 1951).

60. Symonds, Percival M. "Reflections on Observations of Teachers," Journal of Educational Research, XLIII (May 1950), pp. 688-96.

61. Taylor, H. "The Influence of the Teacher on Relative Class Standing," 27th Yearbook of the N. S. S. E., Part II (Bloomington, Ill.: Public School Publishing Co., 1928), pp. 97-100.

62. Thomson, Godfrey H. The Factorial Analysis of Human Ability (New York: Houghton Mifflin Co., 1939), pp. 89-95.

63. Walker, H. M. Elementary Statistical Methods (New York: Henry Holt and Co., 1943), 298 pp.

64. Witty, Paul H. "An Evaluation of Investigations of the Effective Teacher," American Psychologist, III (July 1948), pp. 264-65.

65. Yankey, J. V., and Anderson, P. L. "Review of Literature on Factors Conditioning Teacher Success," Educational Administration and Supervision, XIX (October 1933), pp. 511-520.

66. Buros, Oscar Krisen. The Third Mental Measurements Yearbook (New Brunswick: Rutgers University Press, 1949), 1246 pp.

TEMPERAMENT AND TEACHING SUCCESS

HAROLD WESLEY MONTROSS*
University of Wisconsin Extension
Rhinelander, Wisconsin

SECTION I

INTRODUCTION

A PERSISTENT problem in the field of teacher education is the determination of the personal qualities necessary for success in teaching.

Research in this area has been greatly extended in recent years. An annotated bibliography of teacher competences lists 1006 studies published since 1890 (15). Barr (4) has compiled a summary of 150 of these studies dealing specifically with the measurement and prediction of teaching efficiency.

A survey of the research in the field would seem to show that the teacher's personality is frequently thought important in relation to success in teaching as early as 1910. Ruediger and Strayer (31) found that personality was among the most significant qualities of the successful teachers. Witty (43) analyzed some twelve thousand letters submitted by pupils from all over the nation. The desirable teacher traits mentioned most frequently were: (1) cooperative democratic attitude; (2) kindliness and consideration for the individual; (3) patience; (4) wide variety of interests; (5) general appearance and pleasing manner.

Although research in the field of teacher education would seem to indicate that personality is an important variable in teaching success, the identification and definition of this variable has not been made.

This study is concerned with those aspects of personality sometimes called temperament. Allport (1) has characterized temperament as a certain class of raw material from which personality is fashioned. Cattell (10) describes it as the "constitutional factors" in personality, describing "constitutional" as referring to that part of the personality which, at any age, is least subject to change. Harriman (20) in The New Dictionary of Psychology, defines temperament as:

> The more or less stable effective pattern characteristic of an individual. It is revealed by one's susceptibility to emotional stimulation, by reaction time, by strength of responses, by quality and intensity of moods and by all that is subsumed under 'emotional nature'.

Temperament is used to designate the more or less stable dispositions that are relatively permanent and evidenced at an early age. The earliest adjustments which can be distinguished in infants and which show differences are in the areas of spontaneous activity and emotional expression, —both temperamental traits. Research indicates that these characteristics of early childhood tend to persist (1).

Although temperament might be said to designate a kind of raw material from which personality is fashioned, it probably cannot accurately be said that temperament exists apart from personality, nor is there any personality which is devoid of temperament.

The area of temperament would seem to be one of the logical starting points for a study of the personal prerequisites to teaching success. This would appear so inasmuch as: (1) traits are evidenced at an early age; (2) traits are relatively persistent and stable; (3) traits are said to be the raw material from which personality is fashioned. This investigation is concerned with certain aspects of temperament and their relationship to teaching success.

SECTION II

DESIGN OF THE STUDY

Description of the Subjects

TWO HUNDRED seventy-nine students in the School of Education at the University of Wisconsin secured teaching positions in secondary schools of Wisconsin in the fall of 1950. One hundred one of this group were visited under the beginning teachers visitation program of the University. From this group thirty-five were willing and available to participate in further intensive testing which constituted a part of this investigation.

The thirty-five subjects of this investigation consisted of sixteen male and nineteen female teachers majoring in the areas of English, mathematics, physical education, speech, science, art, music, social science, and home economics.

Each individual of this group had completed a four-year course at the University of Wisconsin and had taught at least one year in the secondary schools of Wisconsin.

---

*The author wishes to express his appreciation to his major professor, Dr. A. S. Barr, for his guidance, encouragement, and inspiration throughout the study.

Table I presents a comparison between the subjects of this investigation and the subjects in a study by Lins (28). Comparisons are made in the following areas: the mean grade point average in Education 73, Education 74, Education 75 practice teaching (required courses in Education at the University of Wisconsin), the total four year grade point average, and the mean percentile rank on the American Council on Education Psychological Examination.

The subjects of this investigation seem not to be greatly different from other groups studied at the University of Wisconsin.

Description of Data-Gathering Devices

A major purpose of the investigation was to study the relationship between certain aspects of temperament and teaching success. Temperament was measured by the Thurstone Temperament Schedule, Cattell's 16 Personality Factor Test, and a series of objective type tests of tempo, fluency, speed, suggestibility, disposition rigidity, and dexterity—coordination.

I. The Thurstone Temperament Schedule (Appendix B):* This is a self-rating test designed to assess those traits which are relatively permanent for each individual. The test was constructed through a factorial study with three hundred and forty questions about personality which seemed effective in differentiating between persons who would be considered well adjusted and which would describe behavior resulting from a relatively stable trait. Thurstone found that seven factors accounted for the variance. These he has tentatively named and described as: (37)

a) Active (A). A person scoring high in this area usually works and moves rapidly. He is restless whenever he has to be quiet. He likes to be "on the go" and tends to hurry. He usually speaks, walks, writes, drives and works rapidly, even when these activities do not demand speed.

b) Vigorous (V). A person with a high score in this area participates in physical sports, work requiring the use of his hands and the use of tools, and outdoor occupations. This group emphasizes physical activity using large muscle groups and great expenditure of energy.

c) Impulsive (I). High scores in this category indicate a happy-go-lucky, daredevil, carefree, acting-on-the-spur-of-the-moment disposition. The person makes decisions quickly, enjoys competition, and changes easily from one task to another. The decision to act or change is quick regardless of whether the person moves slowly or rapidly (Active), or enjoys or dislikes strenuous projects (Vigorous).

d) Dominant (D). People scoring high on this factor think of themselves as leaders, capable of taking initiative and responsibility. They are not domineering, even though they have leadership ability. They enjoy public speaking, or organizing social activities, promoting new projects and persuading others.

e) Stable (E for emotionally stable). Persons who have high Stable scores usually are cheerful and have an even disposition. They can relax in a noisy room, and they remain calm in a crisis. They claim that they can disregard distractions while studying. They are not irritated if interrupted when concentrating, and they do not fret about daily chores.

f) Sociable (S). Persons with high scores in this area enjoy the company of others, make friends easily, and are sympathetic, cooperative, and agreeable in their relations with people.

g) Reflective (R). High scores in this area indicate that a person likes meditative and reflective thinking and enjoys dealing with theoretical rather than practical problems. Self-examination is characteristic of reflective persons. These people are usually quiet, work alone, and enjoy work that requires accuracy and fine detail.

The Schedule was designed to provide a quick survey of the seven areas of temperament. The reliabilities for the seven areas in the Schedule were computed by the split-half method and by the test-retest method. These reliabilities are summarized in Tables II and III.

The tests were administered during the summer and fall of 1951 at the convenience of the subjects. Standard directions were given and the subjects were allowed as much time as necessary to complete the test.

II. The Cattell 16 Factor Personality Test (Appendix C): On the basis of factorial analysis Cattell has designed a questionnaire type test which sets out to measure 16 personality source traits which he feels provide the fundamental structure of personality. The 16 P. F. Test

---

*All references to Appendices may be found in original manuscript on file in Library, University of Wisconsin.

TABLE I

COMPARISON OF THE MEAN GRADE POINT AVERAGE IN EDUCATION 73, 74, 75 PRACTICE TEACHING, TOTAL FOUR YEAR GRADE POINT AVERAGE, AND MEAN PERCENTILE RANK ON THE AMERICAN COUNCIL ON EDUCATION PSYCHOLOGICAL EXAMINATION WITH THE SUBJECTS IN A STUDY BY LINS

Measure	Mean	
	1[1]	2[2]
Education 73	2.31	2.34
Education 74	2.13	2.02
Education 75, Practice Teaching	2.37	2.17
Total Four Year Grade Point Average	1.93	1.88
American Council on Education Psychological Examination	65.63	54.45[3]

[1] Subjects of this investigation
[2] Subjects in a study by Lins
[3] Adjusted mean for men and women

TABLE II

SPLIT-HALF RELIABILITY COEFFICIENTS FOR THE SEVEN AREAS OF TEMPERAMENT

Area	Correlation	
	Men	Women
Active	.48	.46
Vigorous	.61	.63
Impulsive	.65	.65
Dominant	.77	.77
Stable	.63	.64
Sociable	.68	.73
Reflective	.73	.62
N	200	157

TABLE III

TEST-RETEST RELIABILITY COEFFICIENTS FOR THE SEVEN AREAS OF TEMPERAMENT

Area	Correlation
Active	.78
Vigorous	.78
Impulsive	.79
Dominant	.82
Stable	.61
Sociable	.73
Reflective	.75
N	81

TABLE IV

SPLIT-HALF RELIABILITY COEFFICIENTS FOR THE 16 P.F. TEST

Trait	Correlation
A	.84
B	.70
C	.71
E	.82
F	.85
G	.56
H	.74
I	.54
L	.55
M	.72
N	.65
O	.88
Q_1	.50
Q_2	.61
Q_3	.53
Q_4	.76
N	200

was used as an assessment of the total personality and as an indicator of temperamental factors. Cattell says of the test:

> The questionnaire thus aims to leave out no important aspect of the total personality, for the above factors are based on even sampling from the personality sphere and include abilities (intelligence), temperamental factors, and dynamic (character integration) source traits. (7)

The questionnaire has equivalent A and B forms, equal in length and design. Both forms comprising three hundred seventy-four items were used. Each item is followed by the possible response of "Yes", "?", and "No". Cattell has suggested the "forced choice" answer with the elimination of the "?" response under certain conditions. In this investigation the "forced choice" response was used. The tests were administered to the subjects at their convenience in the summer and fall of 1951. Each subject was allowed as much time as necessary to complete both forms of the test.

The split-half reliability of the test on a sample of the general population, corrected to the full number of items in A and B forms is summarized in Table IV.

III. The Objective Type Measures of Temperament (Appendix D): It was felt that tests which demanded no subjective self-evaluation might be most appropriate in this field. The background for the use of such tests has been most fully developed by Cattell (9). He has presented tests which he terms "objective measures" that are thought to indicate or assess certain temperamental factors. Out of this, and other works, tests were devised and adopted. The areas of tempo, fluency, speed, suggestibility, disposition rigidity, and dexterity-coordination were investigated. At least two measures were obtained from each area except for suggestibility where only one measure was used.

The battery of objective type measures required approximately one hour for administration. The tests were administered individually during the summer and fall of 1951 at the convenience of the subjects.

Description of Objective Type Instruments

Test 1: Tempo

Tempo, in this investigation, refers to the natural or preferred rate of speed in any activity.

A number of writers, among them Frischeisen, Kohler, Guttman, Meumann, and Wu have conducted research in the area of tempo and have developed theories of temperament on the basis of the existence of general personal tempo. The most extensive work in the study of tempo itself has been done by Allport and Vernon (2) and by Harrison (21). Allport and Vernon studied fourteen varied natural tempo rates. They were unable to show the existence of a general preferred rate or speed, finding an average intercorrelation between the fourteen rates studied of .17. Harrison also was unable to find a single general factor of tempo. Cattell (7) has postulated that tempo can be clustered into the following areas: (a) ideomotor tempo, evidenced through such activities as the normal rate of reading, counting, handwriting, tapping, and blackboard writing; (b) rhythm tempo evidenced through such activities as the normal rate of finger and hand tapping, leg tapping, patting, counting and stylus compression; (c) drawing tempo, evidenced through such activities as the normal rate of drawing figures on paper, foot drawing and blackboard drawing; (d) general tempo, evidenced through such activities as the normal rate of cranking, tapping, body bending, patting, speed decision, and card sorting; (e) body tempo, evidenced through such activities as the normal rate of head turning, body bending and arm raising; and (f) limb tempo, evidenced through such activities as the normal rate of arm raising and walking.

The existence and the possible temperamental origin of personal tempos seems to be fairly well established in the literature.

The six measures of tempo, used in this investigation, were designed to cover the areas of ideomotor and general tempo. The tests themselves are adaptations of measures generally used in this area.

> Subtest 1a: Subject read a mimeographed paragraph of American history. Subject was told in all tempo tests that time was not a factor in the test, but time was unobtrusively taken by tester.
>
> Score: Time of reading in seconds
>
> Subtest 1b: Subject wrote two lines of the paragraph used in subtest 1a. Time recorded as in subtest 1a.
>
> Score: Time of writing in seconds
>
> Subtest 1c: Subject counted orally to the number forty. Time recorded as in subtest 1a.
>
> Score: Time in seconds
>
> Subtest 1d: Subject sorted cards used in subtest 3b. Time recorded as in

subtest 1a.

Score: Time in seconds

Subtest 1e: Subject tapped on tapping instrument used in subtest 3a. Time recorded as in subtest 1a.

Score: Number of taps in thirty seconds

Subtest 1f: Subject was asked to select from twelve cards face downward on a table, the three cards which he felt were the highest of the twelve. On second trial subject was asked to select the three cards which he felt were the lowest from another set of twelve cards placed in the same position on the table. Time recorded as in subtest 1a.

Score: Average time, in seconds, of two trials

Test 2: Fluency

Fluency of association, referring to the spontaneous calling to mind of ideas, has been shown to exist as a distinct factor. Carroll (5) and Johnson (23), in a factoring of verbal abilities, both found a factor characterized by the ready, spontaneous flow of ideas, words or images. Cattell (7) suggests the presence of at least four distinct kinds of fluency: (a) fluency of association under restriction evidenced through such activities as giving as many words as possible, under time limits, beginning with a certain letter; (b) fluency of association with minimum direction evidenced through such activities as number of words per minute expanding on a theme or topic or number of relevant words per minute describing a picture; (c) fluency and facility in oral speech evidenced through such activities as number of relevant words in describing a picture; and (d) verbal speed evidenced through such activities as speed of handwriting and maximum speed of reading.

It may be noted that Cattell makes a sharp distinction between fluency as a temperamental trait and its association with an ability factor. He says:

> Fluency is one of those temperamental tendencies which also outcrop in abilities and are mistaken for abilities. Thus this same fluency shows itself, apparently, as Thurstone's W, or verbal fluency factor, sharply distinguished from V, or verbal ability, which is itself essentially a knowledge of vocabulary and correct grammatical usage. (10)

Three measures of fluency were used in this investigation. Two of them were developed to assess fluency of association under restrictions. The remaining measure was felt to assess fluency of association under minimum direction.

Subtest 2a: Subject was asked to record on a numbered sheet of paper as many adjectives that he could think of which could be used to describe a house. Time limit four minutes.

Score: Number of relevant words recorded

Subtest 2b: Subject was asked to record on a numbered sheet of paper as many things that he could think of which are 'round' or that are called round. Time limit four minutes.

Score: Number of relevant words recorded

Subtest 2c: Subject was asked to record on a numbered sheet of paper as many words that he could think of which begin with the letter B. Time limit four minutes.

Score: Number of relevant words recorded

Test 3: Speed

Speed in this investigation refers to the speed of repetitive or selective performance where all of the content is perceptually given. Much of the concern in the literature centers about speed as a general or specific factor in abilities. Harrison's work (21) seems to indicate that there is no single general factor of maximum speed. He studied a number of performances such as, card sorting, tapping, head turning, counting, reading, making decisions and found the mean intercorrelations to be about .15. The number of different types of speed performances has not been determined with respect to present research. Cattell (9) feels, though, that from the present research we can assume the existence of six speed factors: (a) speed of judgment and ideomotor performances; (b) speed of reaction time; (c) speed of bodily movement; (d) maximal rhythm; (e) fluency or speed of mental output and association; (f) various factors of natural tempo; and (g) perceptual speed. The relationship of speed and temperament has not been determined although Cattell suggests a general positive correlation. It has been further suggested that the significance for temperament lies in the inter-relationships between speed, tempo and fluency. The research of Studman (36), for example, suggests that the difference between speed and fluency seems to be of special significance. Cattell points out their significance for this investigation as follows:

> The great interest of these speed, tempo, and fluency factors for personality, which justifies the somewhat detailed study of their ability roots here, is that they are among the most direct indicators of temperament (9).

The tests in this area were selected to secure measures from the first two of Cattell's speed factors, those of ideomotor and reaction time speed. These factors represent speed under willed action which is distinguished from tempo and fluency by the definitions used in this investigation.

Subtest 3a: Subject tapped on tapping instrument. The apparatus consisted of a metal plate wired to an electrical counter. A metal tapping pencil was held in the subject's dominant hand. With each tap of the pencil, contact with the metal plate activated the electrical counter. Subject was asked to tap as fast as possible for thirty seconds. Four test readings were taken. There was a thirty second rest period between each test.

Score: Average of four readings

Subtest 3b: Subject was told to sort a set of twenty-four cards into six labeled areas. The set of cards were numbered from two to seven and sorting was by number only. Four test readings were taken. The labeled areas were changed on each test in order to minimize the factor of learning. Subject was asked to go as rapidly as possible.

Score: Average time in seconds of four readings

Subtest 3c: Subject was checked on forty reactions in a visual series with an irregularly spaced interval of warning. The signal used was 'ready', given by the tester before flashing an amber light spaced midway between two keys, one of which had to be depressed to stop the timer. Subject used dominant hand and most convenient key during the test. Twenty reactions were taken; then the Body Sway Suggestibility Test was interjected before checking the final twenty reactions. Timing was in hundredths of a second.

Score: Median reading

Subtest 3d: This test is taken from the Wechsler Test of Mental Ability (41). Subject was asked to write beneath a series of numbers an arbitrarily assigned symbol of each number. Time one minute. Subject was asked to go as rapidly as possible.

Score: Number of digits with correct symbol

Test 4: Suggestibility

Many writers have investigated suggestibility. Among them McDougall, Morgan, Allport, Hull and Eysenck. Eysenck (16) has perhaps presented the most complete investigation of the trait. Out of his factorization of sixteen altogether different tests of suggestibility he postulates the existence of two and possibly three distinct kinds of suggestibility. These he terms 'primary suggestibility,' 'secondary suggestibility,' and 'prestige suggestibility.'

Eysenck further investigates the personality correlates of primary suggestibility in analyzing the scores of one thousand four hundred fifty subjects. He shows significant relationships between his primary suggestibility and traits considered in this investigation to be of temperamental origin.

A well-known measure of suggestibility, and one used by Eysenck, is the Body Sway Test. It was used in this investigation.

Subtest 4a: Subject was blindfolded. To his collar was clipped a string which ran through a pulley taped to the wall and was fastened to a pointer which could move up and down along an upright yardstick, also fastened to the wall. Subject was then asked to stand and relax for one minute and then listen to two amplified recordings. He was told not to try to inhibit or accelerate any feeling or movement, but to try to relax and listen. The forward and backward sway of the subject was then measured on a 1/4 inch sway of the individual.

One minute—the natural sway of the individual.

One-half minute—the sway while the subject listened to the voice recording of "You are slowly falling forward, slowly falling forward; you are slowly falling forward, falling forward, falling forward, falling, falling." Repeat.

One-half minute—the sway while the subject listened to the voice recording of "You are slowly falling backward, slowly falling backward; you are slowly falling backward, falling backward, falling backward, falling, falling." Repeat.

Score: Forward and backward sway of the subject was recorded in inches to the nearest 1/4 inch, positive when it agreed with the direction suggested by the recording, negative when opposed action was shown.

Test 5: Disposition Rigidity

In normal individual behavior it often happens that an old reaction will persist when a new one is intended. In extreme cases, as in the schizophrenic, the individual may sit for long periods of time repeating the same movements or words. This behavior has been termed perseveration by most of the writers in the literature. Cattell (9) in a recent extensive review of the data has postulated that the only assured manifestation of perseveration is as disposition rigidity. In agreement with Cattell this investigation considers disposition rigidity to refer to the resistance of willed change of old established habits. It is best measured by tests in which the individual tries to perform an old task in a new way. For example, it could be measured by the number of times the subject could write his name in one minute divided by the number of times he could write it backward in one minute.

Traditionally, the tendency for behavior to repeat itself when once started, has been associated with temperament. Eysenck (16), though, was unable to find tests of this factor which would distinguish between the groups he studied. Cattell, on the other hand, feels that disposition rigidity is one of the important indices of temperament.

The measures of disposition rigidity in this investigation are taken from a battery of tests devised by Cattell (11) for the assessment of this rigidity trait.

Subtest 5a: Subject was asked, under time limits, to write the digits 2, 3, 7, (X). Time one minute. Subject was asked to write the digits 2, 3, 7, (Y) backwards both in direction and position. Time one minute.

Score: Number of digits written in X, over number of digits written in Y.

Subtest 5b: Subject was asked, under time limits, to write the symbol L (X). Time thirty seconds. Subject was asked, under time limits, to write the symbol ⅃ (Y). Time thirty seconds. Subject was asked, under time limits, to write the two symbols (Z) alternately. Time was one minute.

Score: Total number of symbols in X and Y, over total number of symbols in Z.

Subtest 5c: Subject was asked to trace through a maze (X) laid out in blocks comparable to a city street system. Letters in the maze paths indicated the direction subject was to go. D indicated down, L indicated left, R indicated right and U indicated up. Time one minute. Subject was asked to trace through a similar second maze (Y). In maze Y subject was asked to do the direct opposite of what the letters in the paths suggested. Time was two minutes.

Score: Total number of correct decisions made in X over total number of correct decisions made in Y.

Test 6: Dexterity—Coordination

The significance or temperamental implications of dexterity—coordination is not as clear as in the other objective test measures. Eysenck (16) finds finger dexterity impaired in individuals of certain temperamental type. Cattell (10) has found what he terms some sort, of psychomotor efficiency, related to temperament. Schwartz (35), although not concerned with temperament directly, found a significant correlation (-.60) between a two hand coordination test and field teaching.

On the strength of these suggested relationships two tests were used in this investigation. Both are adaptations by the author which were felt could give the desired assessment for this area.

Subtest 6a: Instrument in Subtest 3a was used. Rather than tap on metal plate subject was asked to put tapping pencil alternately into three holes placed in a triangular arrangement on the metal plate. When tapping pencil touched the bottom of the hole the electrical counter was activated. Time thirty seconds. Subject was asked to go as rapidly as possible. Four trials were given, two with the right or dominant hand and two with the left or non-dominant hand.

Score: Difference between the average

number of taps of right and left hand.

Subtest 6b: A sheet of paper was blocked off into sixty equal squares. With a pencil in each hand subject was asked to mark simultaneously a straight line and a cross within each square successively for thirty seconds.

Score: Number of squares filled with both symbols.

IV. The Criteria of Teaching Success: Teaching success, in this investigation, refers to (a) the acceptability of the teacher by the principal or superintendent; and (b) it refers to teaching effectiveness as that effectiveness is defined on the Wisconsin Adaptation of the M-Blank. (See Appendix A.)

Description of Ratings

Rating 1: Acceptability

A rating, termed "acceptability" by Lamke (26) in a previous investigation, was obtained. To obtain the rating an interview was conducted with the teacher's principal or superintendent, in which, during the course of the interview, the following questions were asked:

1. How do you like this teacher?
2. Does he get along well with students?
3. Do the parents feel the same way as the students do about the teacher?
4. Is the teacher happy here? Does he have friends on the faculty, in the community, and the like?
5. Does the teacher have any particular weaknesses or strengths?
6. In terms of beginning teachers with whom you have worked in the past, and in the light of our discussion as a whole, would you consider this teacher as among the best, above average, average, below average, or failing as a teacher?

Out of his consideration of the questions asked, the principal gave an overall rating of the teacher.

Rating 2: Principal or Superintendent's Efficiency Rating

The rating was given by the teacher's principal or superintendent using the Wisconsin adaptation of the M-Blank as a guide for the assessment of the teacher's effectiveness.

Rating 3: Experienced Rater I

Each subject was visited in his classroom by two experienced raters. A group of six graduate students, trained for the visitation program, formed what was termed the core group. The remaining raters were staff members of the University of Wisconsin. On each visitation at least one of the core group was included. This core group was labeled experienced rater I.

Independent ratings were made of the teacher's effectiveness using the Wisconsin adaptation of the M-Blank as a guide.

Rating 4: Experienced Rater II

This rating was obtained under the same conditions as rating number 3 and was given by the staff members of the University of Wisconsin.

These four ratings were obtained in the teacher's first year of teaching and after he had been teaching for at least three months.

Rating 5: Principal or Superintendent's Efficiency Rating in Second Year of Teaching

An abbreviated form (Appendix A) of the Wisconsin adaptation of the M-Blank was used for this rating. It was obtained in the subjects' second year of teaching, at least three months after the beginning of the school year.

Reliability and Validity of Objective Measures of Temperament

Cronbach (14) suggests the consideration of three types of reliability coefficients, that of coefficients of equivalence, stability, and stability and equivalence. The equivalence aspects of reliability estimates how precisely the test measures an individual sample of behavior at a particular time. It can be determined by the split-half method, by the Kuder-Richardson formula, or by immediate parallel testing. The stability aspect estimates whether or not a behavior at the time of testing is typical behavior at any other time. The stability of a behavior can be estimated by the retest method. The stability and equivalence aspects of reliability gives an estimate of change over a period of time and the fluctuations between two supposedly equivalent tests. The use of a particular type of reliability estimate would depend upon the situation. In the objective measures of temperament the primary interest is in whether or not the particular behavior sampled at the time of testing is typical of that same behavior at some later time. This would mean that the coefficient of stability obtained by a retest after a lapse of time would be considered. It did not seem feasible, at the time of testing, to obtain this reliability estimate.

TABLE V

INTERCORRELATIONS OF CRITERIA RATINGS

(N = 34)

Rating	Correlation 1	2	3	4
Principal's Acceptability	...	...	...	...
Principal's M-Blank	.93	...	...	...
Rater Number 1	.72	.55	...	...
Rater Number 2	.64	.57	.73	...
Principal's M-Blank 2	.50	.53	.39	.38

TABLE VI

RELATIONSHIP BETWEEN THE FOUR RATINGS IN COMPOSITE ONE AND THE FIRST GENERAL FACTOR

Rating	General Factor Loading
Principal's Acceptability	.94
Principal's M-Blank	.87
Rater 1	.85
Rater 2	.84

TABLE VII

RELATIONSHIP BETWEEN THE FIVE RATINGS IN COMPOSITE TWO AND THE FIRST GENERAL FACTOR

Rating	General Factor Loading
Principal's Acceptability	.92
Principal's M-Blank	.87
Rater 1	.83
Rater 2	.81
Principal's M-Blank 2	.68

A number of reliability estimates can be found in the literature which fall under the general classification of mechanical and motor tests. A number of these tests are of the same type used in this investigation.

Woodworth (44), for example, reports the retest reliability of reaction time as being about .90 when given under the same conditions as the test for reaction time in this investigation. Greene (18) reports on the use of a tapping test with a maximum retest correlation of about .91. Several batteries of motor tests are further described by Greene. These batteries include subtests of the kind used in this investigation as objective indicators of temperament. One of these, for example, the Stanford Motor Skills Tests (18) has reported retest reliabilities of .75 to .86 for each item in the battery.

Research studies in the areas of tempo, fluency, suggestibility and disposition rigidity are not as numerous as are the studies in the area of mechanical and motor abilities. Subsequently, reliability estimates are not reported as often.

Perhaps the most comprehensive study has been done in the area of tempo by Allport and Vernon (2). They report retest reliabilities, showing consistency of speed with a single task at different times, in such areas as reading, counting, and handwriting to be .88, .85, and .87, respectively.

Although reliability estimates are reasonably high in the literature, any subsequent investigation in this area must establish the retest reliabilities of the objective measures of temperament.

A test is said to be valid when it measures what it purports to measure. The tests used in this investigation as measures of temperament purport to be valid indicators of temperament. The conclusions may be questioned, but aside from their validity as temperamental indicators, their relationship to teaching success as it is defined in this investigation, whatever they may be measuring, is of interest.

Reliability and Validity of Ratings

No reliability estimates of the five ratings used in the criterion were obtained. The intercorrelations of these criterion ratings can be seen in Table V.

As may be seen the Principals Acceptability and the Principals M-Blank rating show a correlation of .93. These two ratings were obtained in the first year of teaching at different times, with different instruments, and with the same individual giving both ratings. It seems reasonable to assume that had the same instrument been used in both ratings that such a reliability estimate would have been higher than the .93 correlation. Rater one and two, using the same instrument at the same time, correlated .73. It can be seen that the relationship between rater one and rater two, and the relationship between the Principal's Acceptability and the Principal's M-Blank rating, is greater than the relationship of the two principals' ratings with either of the ratings one and two.

The Principal's M-Blank rating number two which was obtained in the second year of teaching with a different instrument, and with the same rater as in the Principal's Acceptability and the Principal's M-Blank rating, shows a relationship with the Principal's M-Blank and Principal's Acceptability of .53 and .50 respectively. It is not known, though, whether the change in rating is the result of a different opinion expressed by the individual giving the rating, by a change in the individual being rated or because of a difference in instruments.

Treatment of the Data

The raw scores of the objective measures of temperament, of the Thurstone Temperament Schedule, and of the Cattell 16 P.F. Test were used in all calculations. The rating scores were treated as composites. Intercorrelations between the various criterion scores are given in Table V. Two composite scores were constructed. The first was a summation of the four ratings obtained in the subjects first year of teaching: the Principal's Acceptability rating, the Principal's M-Blank rating, Rater 1 and 2. The second composite was constructed around a fewer number of cases inasmuch as in the second year of teaching nine subjects were no longer teaching. It consisted of the four ratings which composed composite one plus the addition of a second Principal's M-Blank rating. The criterion scores, hereafter, will be referred to as composite one and composite two.

Further relationships between the individual ratings are given in Tables VI and VII.

Since the correlations between the individual ratings and the general factor are all relatively high the ratings can essentially be considered as measuring the same thing.

Throughout the study the Pearson Product Moment Coefficient of Correlation has been employed to determine the amount of relationship between measures of temperament and the criteria.

To determine the level of significance of a coefficient of correlation tables in Lindquist (27) were used. The tables give significance at the five and one percent levels for samples of various size.

TABLE VIII

CORRELATIONS BETWEEN THURSTONE TEMPERAMENT SCHEDULE AND COMPOSITE ONE AND TWO

Temperament Trait	Composite No. 1	Composite No. 2
A	.01	.05
V	-.23	-.29
I	-.05	-.10
D	.26	.14
E	-.03	-.08
S	.05	-.16
R	.03	.17

TABLE IX

CORRELATIONS BETWEEN THE CATTELL 16 P. F. TEST AND COMPOSITE ONE AND TWO

Trait Trait	Composite No. 1	Composite No. 2
A	.27	.40
B	.11	.18
C	.26	.20
E	.03	.10
F	-.01	.10
G	-.21	-.11
H	-.15	.02
I	.21	.14
L	.06	-.11
M	-.12	-.01
N	.14	.30
O	-.01	-.01
Q_1	.21	.32
Q_2	-.11	.00
Q_3	-.02	-.08
Q_4	-.03	.06
N	35	25

TABLE X

CORRELATIONS BETWEEN THE OBJECTIVE MEASURES OF TEMPERAMENT AND COMPOSITE ONE AND TWO

Measure	Correlation* Composite No. 1	Correlation* Composite No. 2
Tempo Reading	.10	.03
Tempo Writing	.03	.00
Tempo Counting	.06	.23
Tempo Tapping	.10	.14
Tempo Card Sorting	.06	.31
Speed of Decision	.05	.00
Speed Tapping	.45	.47
Speed Card Sorting	.08	.26
Digit Symbol	.17	.36
Reaction Time	.12	.11
Fluency (Adjectives)	.41	.28
Fluency (Things Round)	.09	.03
Fluency (First Letter)	.29	.21
Two Hand Coordination	.13	.21
Right and Left Hand Coordination	.57	.47
Sway (Total Amount)	.02	.20
Disposition Rigidity (Subtest 5a)	.07	.14
Disposition Rigidity (Subtest 5b)	.20	.19
Disposition Rigidity (Subtest 5c)	.11	.08
N	34	26

*All signs changed to indicate positive relationship

TABLE XI

CORRELATIONS BETWEEN COMBINATIONS AND PARTS OF THE OBJECTIVE MEASURES OF TEMPERAMENT AND COMPOSITE ONE AND TWO

Measure	Correlation* Composite No. 1	Correlation* Composite No. 2
Maze Number 1 (Subtest 5c)	.42	.36
Deviation of Reaction Time from Median in Positive Direction	.51	.39
N	34	26

*All signs changed to indicate positive relationship.

SECTION III

ANALYSIS OF THE DATA

Correlations Between Thurstone Temperament Schedule and Composite One and Two

THE CALCULATED correlations for the seven areas of temperament in the Thurstone Temperament Schedule, showing their relationship to composite one and two, are given in Table VIII. The raw scores for the Schedule are given in Table IX. None of the correlations differ greatly from zero.

Correlations Between the Cattell 16 P. F. Test and Composite One and Two

The correlations for this assessment of total personality are shown in Table IX. The raw scores for the 16 P. F. Test are given in Table XVII. None of the correlations except the .40 correlation between Factor A and composite two differ greatly from zero. The correlations between Factors N and Q_1 and composite two were .30 and .32, respectively.

Correlations Between the Objective Measures of Temperament and Composite One and Two

The correlations for the objective measures of temperament and the composites are presented in Table X. Raw scores of the objective measures are shown in Table XVIII.

The correlations of Speed of Tapping (.45), Right and Left Hand Coordination (.57), and Fluency (.41), with composite one are among the larger coefficients of correlation.

The correlations of Speed of Tapping (.47) and Right and Left Hand Coordination (.47) with composite two approximate those with composite one. The correlation between Fluency (.28) and composite two, while positive, is somewhat less.

Correlations Between Combinations and Parts of the Objective Measures of Temperament and Composite One and Two

In Section II it was suggested that the relationship between several of the objective measures might be significant in any assessment of temperament. As, for example, the relationship between speed, tempo, and fluency as possible temperamental indicators. Following this lead several combinations of measures were made. Two of them yielded statistically significant correlations. These are presented in Table XI. The raw scores for these measures are presented in Table XIX.

These correlations are consistent for both composites, except for Maze Number 1 and composite two, where r = .36.

Intercorrelations of Objective Measures Significantly Correlated with Criteria

From an inspection of Tables X and XI, it may be seen that five of the objective measures of temperament show a significant relationship with the criteria, that is, if one is willing to look upon the group here studied as randomly drawn, which strictly speaking, it is not. A summary of these measures is presented in Table XII.

The intercorrelations for the five objective measures showing a significant relationship with the criteria were computed separately for each composite, since the number of cases in each differ.

For composite one the intercorrelations are presented in Table XIII.

A multiple correlation was then computed using the five objective measures of Speed and Tapping, the Deviation of Reaction Time in a Positive Direction, Fluency (number of adjectives describing a house), Right and Left Hand Coordination, and Maze Number 1. Using the Aitkin Modified Method of Pivotal Condensation, as the means of calculation, the value of this multiple correlation was. 734 with an expected shrinkage to .69 (29).

The intercorrelations of the five objective measures for the twenty-six cases which comprise Composite two are presented in Table XIV.

Using the same type of procedure as above a multiple correlation was computed, with a value of .66 and an expected shrinkage to .596.

The multiple correlation coefficient for composite two would ordinarily be expected to be lower than the multiple correlation for composite one since the range in the criterion scores in the second composite is less than in composite one. The standard deviation for composite one is .92, for composite two it is .71. It may be recalled that composite two is the criterion score for those subjects in their second year of teaching and that nine subjects from the original test group were no longer teaching in the second year. Six of these nine subjects were below the general mean of the criterion scores. Accordingly, the subjects of composite two were more homogeneous with respect to the criteria ratings and the objective measures of temperament, in general, were less discriminatory for this more homogeneous group.

Factor A of the Cattell 16 P. F. Test correlated .40 with the composite two criteria. This factor was included in a multiple correlation with the five objective measures. It raised the

TABLE XII

OBJECTIVE MEASURES SIGNIFICANTLY CORRELATED WITH CRITERIA

Measure	Composite No. 1	Composite No. 2
Speed Tapping	.45*	.47**
Deviation of Reaction Time in a Positive Direction	.51*	.39**
Fluency (Adjectives)	.41**	.28
Right and Left Hand Coordination	.57*	.47**
Maze Number 1	.42**	.36
N	34	26

* Significant at 1% level
**Significant at 5% level

TABLE XIII

INTERCORRELATIONS OF THE FIVE OBJECTIVE MEASURES SIGNIFICANTLY CORRELATED WITH CRITERIA (COMPOSITE ONE)

(N = 35)

Measure	Correlation* 1	2	3	4
Speed Tapping	...	...	...	...
Deviation of Reaction Time in a Positive Direction	.15	...	...	...
Fluency (Adjectives)	.19	.08	...	...
Right and Left Hand Coordination	.51	.37	.28	...
Maze Number 1	.35	.37	.28	.32

*Signs changed to indicate positive relationship

TABLE XIV

INTERCORRELATIONS OF THE FIVE OBJECTIVE MEASURES SIGNIFICANTLY CORRELATED WITH CRITERIA (COMPOSITE TWO)

(N = 26)

Measure	Correlation			
	1	2	3	4
Speed Tapping	...	...	...	...
Deviation of Reaction Time in a Positive Direction	.05	...	...	...
Fluency (Adjectives)	.12	.01	...	...
Right and Left Hand Coordination	.48	.21	.20	...
Maze Number 1	.37	.23	.28	.17

TABLE XV

CORRELATION BETWEEN THE FIVE OBJECTIVE MEASURES OF TEMPERAMENT AND FACTOR A OF THE CATTELL 16 P. F. TEST

Objective Measures	Correlation* Factor A
Speed Tapping	.04
Deviation of Reaction Time in a Positive Direction	.04
Fluency (Adjectives)	.20
Right and Left Hand Coordination	.48
Maze Number 1	.07

*All signs changed to indicate positive relationship value of the multiple correlation coefficient to .71.

Although Factor A, when added to the objective measures as a predictive factor, does raise the multiple correlation such a gain is not made without losses.

The relationship of Factor A is presented in Table XV.

Study of Individual Cases

Material for representative case studies was available in the data booklet* used at the University of Wisconsin each year to collect information from students preparing for the teaching profession.

Six case studies were made, three ranked in the first quartile on the criteria scores (Subjects 5, 21, 24) and the remaining three ranked in the fourth quartile (Subjects 4, 28, 35).

It was felt, from the data available, that the following areas would be of importance: sex, marital status, data on size of the family, the parents educational experiences, number and kinds of educational awards, the reasons for entering the teaching profession, any work experience during college, work experience during vacation periods, all extra-curricular and other activities participated in by subject, and the subject's answer to the question of what he would do if he had plenty of money.

In order to make the data more accessable for comparison, the case history of each subject will be presented in graph form on the next few pages.

In reviewing the case studies of those subjects ranked in the first and fourth quartile there does not seem to be any clear-cut distinguishing areas. It may be significant to note, however, that in general those ranked in the upper quartile tended to place teaching near the top in their choice, when asked the question of what they would do if money were no object; those in the lower quartile tended to place teaching near the bottom of the list, or as in two cases used in this investigation, eliminated it entirely.

There does seem to be some evidence to indicate that those teachers ranked in the upper quartile list more extra-curricular activities than do those in the lower quartile.

With the exception of the above areas there did not appear to be any clear cut differences between upper and lower quartile teachers in the data available for the case studies.

SECTION IV

SUMMARY AND CONCLUSIONS

Statement of the Problem

THE PURPOSE of this investigation is to study the relationship between certain aspects of temperament and teaching success.

Design of the Study

The subjects of this investigation were thirty-five students who had graduated from the University of Wisconsin and who had secured teaching positions in secondary schools of Wisconsin in the fall of 1950.

Aspects of temperament were measured by (a) the Thurstone Temperament Schedule, (b) the Cattell 16 Personality Factor Test, (c) a series of objective tests of tempo, fluency, speed, suggestibility, disposition rigidity, and dexterity—coordination.

The criteria of teaching success was developed from five ratings of each teacher: (a) the principal's acceptability, (b) the principal's M-Blank, (c) a rating given by one of a core group of graduate students trained in the visitation program and who visited the teacher in his classroom, (d) a rating given by a staff member of the University of Wisconsin, and (e) a second M-Blank rating given by the principal in the second year of teaching.

Using a composite score for the criteria, and the raw score from the measures of temperament, Pearson product-moment correlations were calculated to determine the relationship between the aspects of temperament and teaching success.

Summary of Findings

No significant relationships were found between the seven areas of temperament of the Thurstone Temperament Schedule and the criteria.

In the Cattell 16 P. F. Test, only one factor, namely A, appeared significant. This was with composite two and was not consistent with correlation of composite one. When placed in a multiple correlation with measures found significant in the objective tests it increased, however, the value of the multiple correlation from .66 to .714.

---

*Teacher Personnel Research Committee, Data Booklet, School of Education, University of Wisconsin, Madison, Wisconsin.

SUBJECT 5

Sex	Marital Status	Family Data	Parents Education	Educational Awards	Reasons for Entering Teaching
Female	Single	Two older brothers	Father: completed the twelfth grade Mother: completed one year of college and is a graduate nurse	None	1. Advised by another person to teach 2. Enjoys working with children 3. Economic security

Work Experience in College	Work Experience during Vacation	Extra Curricular, Other Activities	Answer to Question: If you had plenty of money what would you wish to do?
Library Assistant Student office help Estimates she earned 20% of living expenses while attending college	Girl Scout camp counselor Playground director	Field hockey Softball Games Tennis Radio play Drama Club play Junior class play Declamations Radio Workshop College play Band Glee Club Operetta Musica festivals Feature writer for class annual Camp counselor Pep club Girl Scouts G. A. A.	1. Theater 2. Teaching

SUBJECT 21

Sex	Marital Status	Family Data	Parents Education	Educational Awards	Reasons for Entering Teaching
Male	Married	One younger brother One older sister	Father: Five years of college; M. A degree Mother: Two years of college	Elks Club citizenship essay contest Midwest representative of American Forum of the Air	1. Relatives, friends and parents influenced him 2. Enjoys working with children 3. Economic security

Work Experience in College	Work Experience during Vacation	Extra-Curricular, Other Activities	Answer to Question: If you had plenty of money what would you wish to do?
None Estimate earned 60% of college living expenses (vacation work)	Camp counselor Telephone pay collector Laborer Office work Store clerk	Basketball Football Track Golf Log-rolling Tennis Class play Debate Extemporaneous speaking Mixed chorus Co-editor of fraternity paper Member of National Forensic League National Honor Society Delta Tau Delta Delta Sigma Rho Lyceum Club Hi-Y Club Letter Club College freshman class vice-president Student life committee	1. Some type of travel work 2. Teach debate in college 3. Professional golfer 4. Public Lecturer 5. Trial lawyer

SUBJECT 24

Sex	Marital Status	Family Data	Parents Education	Educational Awards	Reasons for Entering Teaching
Male	Single	Three older brothers	Father: completed 10th grade	Valedictorian of grade school	1. Relatives and friends influenced him
		One older sister	Mother: completed 12th grade	Sophomore honors	2. Feels it is a dynamic profession
		One younger sister			3. Interested in academic subjects

Work Experience in College	Work Experience during Vacation	Extra-Curricular, Other Activities	Answer to Question: If you had plenty of money what would you wish to do?
House Fellow Factory worker Estimates 25% of living expenses earned in college	Department store employee Factory worker	Soccer Volley ball Basketball Tumbling Swimming Ice skating Golf Bowling Class play Oratory Debate Extemporaneous speaking Schreders Orchestra Chorus Solo work Mixed chorus Church choir Phi Eta Sigma Student council	1. Travel and attend all types of schools 2. Meet and study people all over the world 3. Work for cooperative friendship between all groups 4. Write about social problems 5. Teach social problems 6. Teach anything

SUBJECT 4

Sex	Marital Status	Family Data	Parents Education	Educational Awards	Reasons for Entering Teaching
Female	Single	One older brother	Father: completed 8th grade	None	1. Relatives and friends influenced her
		Two younger brothers	Mother: completed 8th grade		2. Admires teachers and their work
		Four younger sisters			3. Feels that teaching is a profession in which she is capable of success

Work Experience in College	Work Experience during Vacation	Extra-Curricular, Other Activities	Answer to Question: If you had plenty of money what would you wish to do?
Housework	Billing clerk	Tennis Softball	1. Architecture
Waitress	U. S. Navy	Swimming Church choir	2. Interior decorating
Salad girl	Factory laborer	Reporter for school paper	3. Book critic
Librarian			
States she earned all her living expenses in college			

SUBJECT 35

Sex	Marital Status	Family Data	Parents Education	Educational Awards	Reasons for Entering Teaching
Female	Single	Only child	Father: five years of college; M.S. Degree Mother: five years of college; M.S. degree	High school valedictorian	1. Relatives and friends influenced her 2. Economic security

Work Experience in College	Work Experience during Vacation	Extra-Curricular, Other Activities	Answer to Question: If you had plenty of money what would you wish to do?
Director of arts and crafts in YWCA	Factory worker	Class play Wisconsin players Art club Dramatic club Craft work Annual staff Art editor of newspaper French club Girl Scouts Vice-president of high school class	1. Artist-craftsman 2. Literary critic 3. Homemaker 4. Professional writer 5. Musician 6. Illustrator

SUBJECT 35

Sex	Marital Status	Family Data	Parents Education	Educational Awards	Reasons for Entering Teaching
Female	Single	Two older brothers	Father: four years of college	National Honor Society	1. Advised by parents and friends
			Mother: completed two years of college	Quill and Scroll	2. Economic security
					3. Enjoys working with children

Work Experience in College	Work Experience during Vacation	Extra-Curricular, Other Activities	Answer to Question: If you had plenty of money what would you wish to do?
Office help	None	Basketball Volleyball Tennis Archery Swimming Plays Reporter on school paper Mixed chorus Choir Operettas Latin club War Stamp club Press Club Girl Reserves Junior Red Cross	No comment

Five of the objective measures of temperament were found to be significantly correlated with the criteria, namely, (a) Speed of Tapping, (b) Deviation of the Reaction Time in a Positive Direction from the Median, (c) Fluency (adjectives), (d) Right and Left Hand Coordination, and (e) Maze Number 1. These correlations ranged from .39 to .57.

A multiple correlation of .734 was secured for composite one, and of .66 for composite two when these five objective measures were employed in a regression equation.

Conclusions

In light of the results of this investigation it seems reasonable to postulate that there may be certain temperamental patterns which will distinguish between good and poor teachers as measured by ratings of principals and others trained to evaluate teacher effectiveness.

It would seem from the data in this investigation that identification of these temperamental patterns can best be accomplished through the use of objective measures. The Thurstone Temperament Schedule and the Cattell 16 P.F. Test seemingly fail to identify aspects of temperamental behavior which are related to success in teaching as measured in this investigation.

It may be argued that the objective measures used in this investigation are restricted to measuring only what they literally measure in the particular test situation, that is, the Speed of Tapping Test measuring the speed of the subjects tapping in the particular situation in which the sample of behavior was observed. It seems, however, that further and broader interpretations might be inferred from these test situations.

If one makes these interpretations a "good" teacher may be described as follows: The "good" teacher would seem to be characterized as possessing a particular type of fluency. He has the ability to associate ideas and things. He would seem to be able to maintain a definite mental set, that is, he can better concentrate, over a period of time, upon an assigned task, as evidenced through the small variation between his median score on the Reaction Time Test and the deviation of his reactions in a positive direction from his own median. Relatively, he would seem to show more interest and concern for the task at hand and seems to have a desire to work near the upper limits of his ability most of the time in accomplishing that task. In somewhat the same vein he may be characterized as having more of what can be termed "drive" or determination to succeed. Finally, the "good" teacher would seem to possess more speed in performing tasks of the psychomotor type. Although this would seem to be of temperamental origin its temperamental correlates in observable behavior are not clear.

BIBLIOGRAPHY

1. Allport, G. W. Personality: A Psychological Interpretation (New York: Henry Holt and Co., 1937).
2. Allport, G. W., and Vernon, P. E. Studies in Expressive Movement (New York: The Macmillan Co., 1933).
3. American Association for Health, Physical Education and Recreation, Research Methods Applied to Health, Physical Education and Recreation (Washington, D.C., 1949).
4. Barr, A. S. "The Measurement and Prediction of Teaching Efficiency: A Summary of Investigations," Journal of Experimental Education, XVI (June 1948), p. 16.
5. Carroll, J. B. "A Factor Analysis of Verbal Abilities," Psychometrika, VI (1941), pp. 279-308.
6. Cattell, Raymond B. "Temperament Tests, II. Tests," British Journal of Psychology, XXIV (July 1933), pp. 20-49.
7. ——— Handbook for the Sixteen Personality Factor Questionnaire (Champaign, Ill.: Institute for Personality and Ability Testing).
8. ——— Guide to Principles and Techniques of Personality and Ability Testing (Champaign, Ill.: Institute of Personality and Ability Testing).
9. ——— Description and Measurement of Personality (Yonkers-on-Hudson, N.Y.: World Book Co., 1946).
10. ——— Personality: A Systematic Theoretical and Factual Study (New York: McGraw-Hill Book Co., 1950).
11. ——— A Guide to Mental Testing, II (London: University of London Press, 1948).
12. ——— "Confirmation and Clarification of Primary Personality Factors," Psychometrika, XII (1947), pp. 197-220.
13. ——— "The Riddle of Perseveration," Journal of Personality, XIV (1946), pp. 229-267.
14. Cronbach, Lee J. Essentials of Psychological Testing (New York: Harper and Brothers, 1949).
15. Domas, Simeon J., and Tiedeman, David V. "Teacher Competence: An Annotated Bibliography," Journal of Experimental Education, XIX (December 1950), pp. 101-218.
16. Eysenck, H. J. Dimensions of Personality (London: Kegan, Paul, Trench, Trudner and Co., 1947).
17. Foley, J. P. "Factors Conditioning Motor Speed and Tempo," Psychological Bulle-

tin, XXXIV (1937), pp. 315-394.

18. Greene, Edward B. Measurements of Human Behavior (New York: Odyssey Press, 1952).

19. Grieder, Calvin, and Newburn, Harry K. "Temperament in Prospective Teachers," Journal of Educational Research XXXV (May 1942), pp. 683-693.

20. Harriman, Philip L. The New Dictionary of Psychology (New York: Philosophical Library, 1947).

21. Harrison, R. "Personal Tempo and the Interrelationships of Voluntary and Maximal Rates of Movement," Journal of General Psychology, XXIV (1941), pp. 343-379.

22. Humm, Doncaster G. "The Rationale of Temperament Testing," Educational and Psychological Measurement, No. 5 (1945), pp. 383-393.

23. Johnson, D. M., and Reynolds, F. "A Factor Analysis of Verbal Ability," Psychological Record, IV (1941), pp. 183-195.

24. Kemp, L. C. D. "The Prediction of Teaching Success," The Forum of Education (Sidney, Australia: Teachers College, August 1947), p. 14.

25. Kennedy, M. "Speed as a Personality Trait," Journal of Social Psychology, I (1930), pp. 286-299.

26. Lamke, Tom Arthur. Personality and Teaching Success, Ph. D. Thesis, University of Wisconsin.

27. Lindquist, E. F. Statistical Analysis in Educational Research (New York: Houghton Mifflin Co., 1940).

28. Lins, Leo Joseph. "The Prediction of Teaching Efficiency," Journal of Experimental Education, XV (September 1946), pp. 2-60.

29. McNemar, Quinn. Psychological Statistics (New York: John Wiley and Sons, Inc., 1949).

30. Retan, George A. "Emotional Instability and Teaching Success," Journal of Educational Research, XXXVII (October 1943), pp. 135-141.

31. Ruediger, W. C. and Strayer, C. D. "The Qualities of Merit in Teachers," Journal of Educational Psychology, I (1910), pp. 272-278.

32. Seagoe, M. V. "Prediction of In-Service Success in Teaching," Journal of Educational Research, XXXIX (May 1946), pp. 658-663.

33. ——— "Prognostic Tests and Teaching Success," Journal of Educational Research, XXXVIII (May 1945), pp. 685-690.

34. ——— "Standardized Tests in the Pre-Training Selection of Teachers," Journal of Educational Research, XXXVI (May 1945), pp. 678-693.

35. Schwartz, Anthony N. "A Study of the Discriminating Efficiency of Certain Tests of the Primary Source Personality Traits of Teachers," Journal of Experimental Education, XIX (September 1950), pp. 63-93.

36. Studman, L. G. "Studies in Experimental Psychiatry: 5. W and F Factors in Relation to Traits of Personality," Journal of Mental Science (1935), pp. 107-137.

37. Thurstone, L. L. Examiner Manual for the Thurstone Temperament Schedule (Chicago: Science Research Associates, Inc., 1950).

38. ——— "The Factorial Description of Temperament," Science, III (April 28, 1950), pp. 254-255.

39. Traxler, Arthur E. "Measurement in the Field of Personality," Education, LXVI (March 1946), pp. 424-430.

40. Vernon, Philip. The Structure of Human Abilities (New York: John Wiley and Sons, Inc., 1950).

41. Wechsler, David. The Measurement of Adult Intelligence (Baltimore: Williams and Wilkins, 1944), 3rd ed., p. 94.

42. Wittenborn, J. R. "Mechanical Ability, Its Nature and Measurement. I. An Analysis of the Variables Employed in the Preliminary Minnesota Experiment," Educational and Psychological Measurement, No. 5 (1945), pp. 241-260.

43. Witty, Paul A. "Teacher Who Has Helped Me Most: Summary of Ratings by Children All Over the Nation," NEA Journal, XXXVI (1947), p. 386.

44. Woodworth, Robert S. Experimental Psychology (New York: Henry Holt and Co., 1938), p. 337.

Journal of Experimental Education

Volume XXIII December, 1954 Number 2

SOCIAL COMPETENCE AND SUCCESS IN TEACHING

ARTHUR SINGER, Jr.*
Northern Illinois State Teachers College
DeKalb, Illinois

SECTION I

STATEMENT OF THE PROBLEM AND HYPOTHESIS

THE TEACHER influences, as a director of learning and according to his abilities, pupil progress toward intelligent self-direction. Some teachers achieve eminent success in their role and by their performance of duties while others accept mediocrity, or even labor in vain with subsequent failure. Success or failure may come as a result of any number of combinations of factors resident in the social and physical environment, in the teacher and in the interaction resultant in the juncture of these two. The kind of academic preparation the teacher has, his personality, his social competence and understanding, the kind of community in which he teaches, the type and condition of the physical plant, the character of the administration, or the kinds of pupils he encounters are only a few of the, literally, hundreds of factors that may influence his efficiency as a teacher. Out of this complex maze of influencing factors, our concern is with the social competency of the teacher.

More specifically, the purpose of this study is to determine, as far as it is possible to do so in a single limited investigation, the relationship between success in teaching and certain aspects of social competence, arising out of inter-personal-group modes.

If the hypothesis that is held is verified, that is, that there is a dynamic relationship between success in teaching and certain aspects of social competence, interaction and inter-personal-group modes, then additional instruments can be used as criteria for further attempts to solve the problem of what constitutes teaching effectiveness.

The investigation of what constitutes success in teaching and the relationship between it and certain aspects of social interaction and inter-personal-group modes will necessarily focus attention on certain questions:

1. Does the concurrent use of instruments such as sociometric techniques, audio-recording ratings, observation, personal adjustment inventories, and teacher ratings provide worthwhile measures of teaching success?
2. Can definite patterns of group social structures be established for the teacher and his classes, and for the teacher and the particular school staff?
3. Is there a relationship between these group social structures and success in teaching?
4. Is there a relationship between the teacher's self-concept of social adjustment and his teaching success?
5. Are the teaching staff and student groups' concepts of the teacher's social acceptability related to efficiency in teaching?
6. If there exists a high or low degree of social acceptability by the group/s for the individual teacher does there also exist a high or low degree of personally judged social adjustment and success in teaching?
7. Does there exist a relationship between classroom oral social competency and success in teaching?
8. Do there exist mean differences between the most successful teachers and the least successful teachers, as evaluated by the various criteria of teaching success, in the measures of social competence?
9. Is the individual teacher who is judged as either most or least successful in teaching also rated correspondingly high or low in so-

---

* The author is deeply indebted to A. S. Barr, Professor of Education, to whom this investigation is dedicated, to Glen Eye, Professor of Education, and to Herbert J. Klausmeier, Assistant Professor of Education, all of the University of Wisconsin.

eau Ednl. Psy. Rese
VID HARE T AI ING C
18/3/55

cial competence as found in the modified sociograms?
10. Is the variable of length of service in teaching a significant factor in the relationship between social acceptance and success in teaching?

Terms Used in this Study

Certain terms are defined below that will be used in the discussion of this investigation. Most all of these terms derive their meanings from the contexts of sociology, psychology, and education. Some of the words have been especially defined for the study. The definitions of terms as they will be construed in this investigation are:

1. Social competence is the human organism's cognition of and adequacy with the mutual relations of mankind living in organized groups or society. This emphasizes the inter and intrahuman psychological function, and total social environment. It includes all the social sub-factors in the individual alone, or in his relationship to other human individuals, such as adjustment, acceptance, interaction, leadership, rejection, group dynamics, introversion, status, etc.
2. Success in teaching for this study means the degree of teacher achieved performance or endeavor in relation to accepted criteria. The terms teacher efficiency and teacher effectiveness will be used interchangeably and mean the same as teacher success.
3. Social interaction is understood as the behavioral, intellectual and verbal processes of human individuals living life in groups, perceiving and cognitive of each other, and with a reciprocal set of relations to one another. These reciprocal relations are not necessarily equal in intensity.
4. Group is defined and refers to two or more people who bear explicit psychological relationship to one another. In this sense it does not merely mean classes of people such as teachers, farmers, or Democrats who are characterized only by some similar property. There may be sub-groups within groups.
5. Dynamic is to be construed as "adjustive changes occurring in group structure as a whole and as produced by changes in any part of the group." (Krech and Crutchfield, 28)
6. The term mode in this study is defined as the natural disposition or manner of existence or action of structure or form.

Importance of the Problem and Further Questions

This investigation attempts a relatively new sort of evaluation of the teacher whereby the educationalist is enabled to obtain a more adequate over-all picture of him in his professional and social relationships. The study will attempt to employ relatively new instruments and techniques in the field of human relationships and social competency which may be relevant to the effectiveness of the teacher. This would assume that knowledge of how the teacher socially conceives of himself and his role in the school as compared to how colleagues, students and others appraise him in that social role is inexplicably interwoven with teaching success.

The proposed investigation is exploratory and does not try to supplant teacher efficiency research up to the present time but rather is only intended as a supplement to it. However, the writer feels it necessary to raise a question in light of completed work in the field to the present. The basic question at issue is just what constitutes an adequate measure of teacher efficiency. The investigator wonders whether teacher self-evaluations, student ratings, expert ratings, and the relationships found between them are adequate measures of teaching or whether these techniques used in conjunction with modified sociometric techniques, audio-recording ratings of classroom teaching and self-rated social introversion and competency measures taken from a personality inventory might not yield a better over-all view of the teacher.

The need for valid, reliable measures of teacher efficiency, so as to more effectively train prospective teachers, improve the in-service training of teachers, and be more unbiased in the promotion and dismissal of teachers, was voiced by Boyce (14) as early as 1915. "In the absence of a valid and reliable criterion for teaching success, solutions for urgent problems must be sought in the realm of expert opinion... one of the crucial problems in evaluation of the education of teachers is the discovery of a satisfactory criterion of teaching success," was stated in 1933 by G. L. Betts (10). In a summary of investigations by Barr (9), we are told that among the results most frequently reported is the need for development of the criteria against which various measures of teaching efficiency may be validated. As late as 1952 Anderson (1) states, "In general no adequate basis for validation of teacher evaluations exists at present. There is apparently no general agreement as to what is good teaching, and even if there were, present day measures lack the reliability necessary for valid criteria." On the basis of possibly adding other criteria for measuring teacher effectiveness, this study may have some importance.

Should positive relationships between various measures employed in this study and teaching success be found, it might be implied that future consideration should be given by teacher institutions, in-service training programs, administrators, and supervisors to the area of social factors. Further, attention should be given to the available techniques in that area for ascertaining the social status of teachers and students preparing for teaching in order to more accurately predict teaching success.

SECTION II

THE METHOD OF INVESTIGATION

THIS IS A descriptive investigation, employing the normative-survey method of study. The subjects, forty-one in-service teachers in eighteen Wisconsin senior high schools, are graduates from the University of Wisconsin between the years 1940-1952, inclusive. Nineteen instruct in various academic subjects and seventeen teach special subject fields; five instruct in both academic and special subjects concurrently (see Table I). All of the teachers are located in schools within a defined geographical area bounded by a circle whose radius extends fifty miles from Madison, Wisconsin.

The Subjects and Locale of the Investigation

The first visit to each school was made to obtain permission from the administration to carry out the investigation. A letter of introduction was carried by the investigator which briefly explained the intended project in the particular school and which was personally addressed to the administrative official in charge. Along with this, representative samples of the testing materials to be used were also shown and explained. Usually the superintendent or principal granted permission immediately, sometimes it was granted as a result of action by the teaching staff after the investigator appeared and explained the project at a teachers' meeting, and in one school permission was granted by the board of education. Such permission, in all cases, was contingent upon the willingness to cooperate on the part of the teachers selected for the investigation. Gracious consent to cooperate was obtained from the forty-one persons during the first interview.

The data which were gathered were of a confidential nature. Cooperation was offered by the schools and teachers with the understanding that it remain such. The schools have all accordingly been assigned an alphabetical symbol and every teacher identified by a number.

Table II shows the group used in this investigation and the subjects used in three other studies made of School of Education graduates from the University of Wisconsin during the years 1940-1952.

Table III contains a year by year comparison, for the years 1940-1952, between the subjects studied and all of the graduating seniors from the University of Wisconsin with respect to their four year over-all grade point averages. It further shows the group studied and graduating seniors from the School of Education, year by year for the thirteen year period, in a comparison of their four year over-all grade point averages.

The teachers studied in this investigation would appear to be representative of the population of the School of Education seniors graduating from the University of Wisconsin with respect to grade point average. In addition, all of the teachers in this investigation appear homogeneous to the larger population composed of School of Education graduates in the following ways: (1) requirements for entrance to the University and School of Education have been met by them, (2) their indoctrination and training experience has been generally similar to that of other School of Education graduates, and (3) upon graduation they have secured state high school positions. While the group may be heterogeneous in many aspects, this also may be said of the parent population.

The Instruments and Procedures Employed in Gathering the Data

The measuring instruments employed in this study may be divided into two groups; those which measure professional teaching competence and those which measure social competence. The instruments used to seek measures of teaching effectiveness were the following: (1) administrator's M-Blank rating, (2) qualified person's M-Blank rating, (3) teacher self-evaluation form, (4) pupil reaction to instruction blank, (5) confidential X-ratings of the teacher by an outside agency, and (6) trained persons' ratings of audio-recordings of teaching presentations. In ascertaining the social competence of the teacher, the data-gathering devices used were as follows: (1) teacher-teacher social distance scale, (2) teacher-student social distance scale, (3) student-teacher social distance scale, (4) Minnesota Multiphasic Personality Inventory, and (5) trained persons' ratings of recordings of teaching presentations. Individual explanation is offered for all of the instruments used. First, the teacher criteria are considered as a group. Explanation is given of data-gathering devices, the procedures used in gathering the data and the reliability and validity of the criteria. Second, and also considered as a group, there follows an explanation of the instruments used to measure social factors, the procedures employed in data-gathering and the validity and reliability of the instruments.

TABLE I

IDENTIFICATION NUMBER OF EACH TEACHER, WITH DATE OF GRADUATION FROM THE UNIVERSITY OF WISCONSIN, AND THE SCHOOL SUBJECTS TAUGHT

	Identification Number	Date of Graduation	English	U. S. History	Mathematics	General Science	Spanish	Latin	Biology	World History	Civics	Speech	Literature	Physics	Chemistry	Social Studies	Voc. Agriculture	Phys. Education	Art	Music	Home Economics	Bus. Education
A	(30	1940	X														X					
	(31	1950											X									
B	(41	1948				X												X				
C	(42	1949				X			X									X				
D	(18	1949				X			X					X	X							
	(19	1948			X																	
	(22	1951																	X			
	(61	1952																			X	
	(26	1951															X					
E	(59	1948	X	X																		
	(52	1951				X												X				
F	(60	1951																		X		
	(2	1949															X					
G	(33	1949		X							X					X						
	(24	1950																X				
H	(45	1949	X																			
I	(11	1948																X				
	(14	1948																X				
	(29	1948	X									X										
	(44	1948		X							X					X						
J	(9	1951		X							X											
	(7	1948		X														X				
K	(48	1950	X				X															
	(57	1949		X												X						
L	(32	1948															X					
	(12	1952	X					X														
M	(20	1951	X																	X		
	(27	1950										X					X					
N	(49	1952																			X	
	(47	1952		X							X											
	(5	1951														X						
	(55	1943	X																			
O	(58	1950		X			X				X											
	(51	1950																		X		
P	(13	1950								X	X					X						
Q	(62	1949																X				
	(53	1950																				X
R	(4	1951	X				X															
	(34	1942							X						X							
	(36	1940																		X		
	(40	1945																X				
	Totals		9	8	1	4	3	1	3	1	6	2	1	1	2	5	5	9	1	4	2	1

Grand Totals: ~~eaching Academic Subjects = 47
Teaching Special Subjects = 22

Note: Some teachers instruct in both academic subjects and in special subjects concurrently, or others teach more than one academic subject. Hence, the total of subjects taught is greater than the total number of teachers.

TABLE III

A COMPARISON OF FOUR YEAR OVER-ALL GRADE POINT AVERAGES OF TEACHERS STUDIED IN THIS INVESTIGATION WITH EDUCATION SENIORS AND ALL UNIVERSITY SENIORS FOR THE ACADEMIC SCHOOL YEARS 1940-52

Academic Year	Teachers in this Study	Four Yr. Grade Pt. Ave.	No. of Educ. Seniors	Four Yr. Over-All Grade Pt. Ave.	No. of University Seniors	Four Yr. Over-All Grade Pt. Ave.
1939-40	2	1.73	341	2.01	3742	1.82
1940-41	0	....	352	2.04	3882	1.85
1941-42	1	2.52	339	2.07	3568	1.87
1942-43	1	1.26	225	2.04	2766	1.90
1943-44	0	....				
1944-45	1	1.57				
1945-46	0	....				
1946-47	0	....	298	1.74	5269	1.91
1947-48	9	1.99	420	1.98	6221	1.90
1948-49	7	1.79	969	2.01	8620	1.88
1949-50	8	2.00	1181	2.02	9188	1.88
1950-51	8	2.26	938	2.02	6615	1.88
1951-52	4	1.64	761	2.05	5678	1.87
Totals and Mean Aves.	41	1.86	5824	1.99	55549	1.87

Note: No records of grade point averages were kept by the University during the war years 1943-46.

TABLE IV

INTERCORRELATIONS OF TEACHING CRITERIA RATINGS

Rating	Correlation 1	2	3
Composite of Supervisory Ratings	.....	.....	.....
Teacher Self-Evaluation	+.079	.....	.....
Pupil Reaction to Instruction	+.565	+.262	.....
Audio-Recording Ratings	+.524	+.171	+.346

The Instruments Measuring Teaching Effectiveness

Administrator's M-Blank Rating Scale:

A one-page modified form of the Wisconsin Adaptation of the M-Blank was used by the investigator. The M-Blank requests the rater to estimate the teacher's strengths and weaknesses in open-ended categories and then rate them in a "General Evaluation" category. During the first visitation to each school where one or more of the teachers selected for study taught, the ranking school official was requested to rate each teacher on this form. In most cases, the superintendent or principal filled in the estimates; in a few cases the assistant principal or school supervisor did it. The score used in this investigation was the "General Evaluation" rating which was divided into categories on a five-point scale as follows: (1) outstanding, (2) above average, (3) average, (4) below average, and (5) poor. The lower the point score the higher was the estimate of the teacher's over-all effectiveness.

Investigator's M- Blank Rating:

Each teacher in the study was rated by the investigator during the time of visitation to the teacher's classroom when the audio-recording was made of one complete teaching presentation. The modified form of the Wisconsin Adaptation of the M-Blank was used for this purpose and the same categories, taken from the "General Evaluation," on a five-point scale were then utilized as scores. A low point score indicates a high rating.

Teacher Self-Evaluation Rating:

Evaluation forms were developed by Harold M. Anderson of the University of Wisconsin for use by pupils, peers and the teacher. The Teacher's Self Evaluation Form of this was placed in the hands of each teacher studied at the time of the first interview. The score used in this study was the over-all evaluation (item 10). The categories of the over-all evaluation were as follows: (1) best, (2) second, (3) middle, (4) fourth, and (5) lowest. A low score indicates a high rating.

Pupil Reaction to Instruction:

Estimates of each of the teachers in the investigation were secured from five students by using the Pupil Reaction to Instruction form developed by Harold M. Anderson. These students were selected from each teacher's classes and the students returned their completed forms in a sealed envelope furnished by the Teacher Personnel Research Committee, directly to the investigator. The teacher was requested to choose one student, the principal chose two students and the investigator chose two students at random to do the ratings. The sum of these five ratings on the last general category, with ratings, (1) best fifth, (2) second fifth, (3) middle fifth, (4) fourth fifth, and (5) poorest fifth, was the source for the pupil evaluation score used in this investigation. A low point score indicates a high rating.

The Confidential X-Ratings:

Little information, other than the method of scoring, can be given here about the X-ratings which were obtained from the Wisconsin State Department of Public Instruction of each teacher in the study because of their confidential nature. The highest possible score that could have been attained was three. The lower the score the higher the rating.

Professional Competence Ratings Based Upon Audio-Recordings of Teaching:

An audio-recording was made of a complete class teaching period of each person in the investigation. During the initial interview a request was made that three different classes and times which the teacher taught and was willing to have recorded be submitted to the investigator. Selection of the particular class to be recorded was then made by the investigator on the basis of the following: (1) recordings were to be made of classes in the teacher's major field where it was felt he would do his strongest teaching, (2) recordings were made in the morning whenever possible so as to insure freshness and alertness on the part of the teacher and students, and (3) availability of time was considered in recording dates so as to have participants relaxed and amenable to the task.

Subsequent to the audio-recordings a rating scale was constructed. It was divided primarily into two sections; one section which was devoted to items measuring professional teaching competence and the other composed of items measuring social competence. Two raters from the staff of the School of Education at the University of Wisconsin concurrently listened to each one of the recordings a minimum of one time and individually rated each teacher for teaching competence. The over-all score for each rater's estimates was obtained by summing the individual items in their five-point scales and then dividing by the number of items. The two rater's average mean scores of professional teaching competence were then combined and divided by two to obtain a single score ranging between one and five. The lower the score the higher is the indicated rating.

Reliability and Validity of Teacher Efficiency Criteria

The only reliability estimate obtained of the six teaching efficiency ratings used was made by ascertaining the reliability of ratings between the two persons who estimated the professional competence of teachers in this investigation from the audio-recordings. It was found that there exists a high coefficient of correlation (+.95) between the over-all ratings of judges one and two on the professional teaching competence scores. This would appear to indicate that for both the items measuring teacher effectiveness in the rating scale and the estimates by the qualified persons, there existed a high degree of agreement. The intercorrelation of the composite score of the three supervisory ratings compared with the other three criterion scores can be seen in Table IV.

It can be observed that there is little correlation between the composite of supervisory ratings and the teacher's self-evaluation of his work. In all cases, the teacher's self-evaluation shows little relationship with either the composite of ratings made by supervisors, pupil estimates of instruction, or ratings of audio-recordings (+.079, +.262 and +.171 respectively). The composite of supervisory ratings does show a significant relationship (+.565 and +.524 respectively) with pupil reactions to instruction and audio-recording ratings. It is not known whether such discrepancies in outside raters is the result of different opinions expressed by the individuals making the estimates, by changes in the teachers rated or by differences inherent in the measuring instruments. The correlation (+.346) was found between pupil reaction to instruction and the audio-recording ratings. Possibly the students were in the position to rate their teachers on the basis of both seeing and hearing their teacher over a much longer period of time, thus the discrepancy between the two ratings.

The Instruments Measuring Social Factors

Before proceeding to a more specific description of the particular instruments used to obtain measures of social factors the investigator would like to explain that much use is made of sociometric techniques in the study. Sociometry is the study of the interrelation patterns existing between people. It is concerned with the actual psychological constituents which motivate and give direction to the function of people's interactive relationships rather than official relationships of office or station in life. Sociometry conceives of the interactive process, with the differing levels of intensity of feelings of attraction, and rejection, as a state of dependency on two or more persons forming social groups. Jennings (25), states "Sociometry aims to uncover the psychological geography of society, the dynamic interpersonal structures which are the base for all social phenomena. . . . the structure in which the latter are imbedded and carried. It contends that to study the latter apart from their locus is futile."

The data-gathering devices employed to measure the social introversion, social competence, social acceptance or distance, and social interaction (friendship choices) of the subjects in relation to the school groups of which they were a part are now considered individually. The three ways in which data were gathered regarding social competence were, first, by employing the Minnesota Multiphasic Personality Inventory from which Social Introversion (SI) and Social Competence (SC) scores were obtained; second, through audio-recordings taken of one of each teacher's classroom teaching presentation and from which ratings were obtained or oral factors of classroom social competence, and third, administering social distance scales to the teachers, staffs and students whereby ratings of social acceptance together with choices for friendship and social competence were secured. A critical discussion of the instruments used, the procedures employed and the reliability and validity of the tests is offered shortly. However, before proceeding further it would be well at this time to again emphasize certain terms and clarify the usage of one of the instruments so as to avoid possible confusion later.

In section one, the term social competence is defined to mean the sum total of social factors as they apply to this investigation. The term is also used to identify certain data resulting from the administration of the Minnesota Multiphasic Personality Inventory, the various social distance scales and ratings of classroom audio-recordings. The original definition of social competence would include all other instances where usage of the same term applies to subparts of the study. Referral to the proper source will be made henceforth wherever the term social competence appears in the investigation.

Teacher-Teacher Social Distance Scale

A social distance scale is used primarily for the purpose of discovering the degree to which individuals accept others in the group/s to which they belong, and how well the group/s accept them. It provides opportunity for each member in a group to react to every other member including reaction to himself regarding how he thinks other members accept him. This provides an extension to the usual sociogram where only a limited number of choices of other group members can be made by each individual.

The instrument used to obtain social ratings by colleagues of each teacher studied, and also to obtain his social rating of every other teaching staff member was adapted and modified for specific use from the Classroom Social Distance Scale originally shown in the booklet, How to Construct a Sociogram by the Horace Mann-Lincoln Institute of School Experimentation, Columbia University (44). The Teacher-Student and Student-Teacher Social Distance Scales which are described later were also adaptations and modifications from the same source. The Teacher-Teacher Social Distance Scale was designed primarily to obtain only social distance ratings from each teaching colleague. Because the administration had status position different from other staff members which might have served to inhibit valid responses, it was decided that the school administration should not have opportunity to participate in such ratings.

Page one of the instrument contains directions for use of the scale while page two includes, for each school, the staff members' names and the five categories of social distance and acceptance which served as a basis for obtaining acceptance scores used in the investigation. Scores between one and five were utilized with a low number indicating the higher score. Page three contains, among other items, two over-all requests which are as follows:

> Would you please write down, on the basis of the five reasons (norms, environment, activities, interaction, and sentiment) listed and explained above, and any you may have added, the three teachers in your opinion who are best in all of these things. Please place them in order of importance.
>
> 1.________ 2.________ 3.________
>
> and
>
> Are the three colleagues you have listed also the three teachers you consider your best friends in school? _____ If not, would you list them in order.
>
> 1.________ 2.________ 3.________

The responses to the above two requests were used to make the modified sociograms of friendship choices and social competence which are shown on pages 112 through 121.

A copy of the Teacher-Teacher Social Distance Scale was placed in the hands of nearly every teaching staff member in each of the eighteen schools together with a letter explaining the investigation. Along with these, an envelope was given to the teachers which was addressed to the Teacher Personnel Research Committee; they had the option of returning the social distance scale to the Committee by mail or by leaving it sealed with the central office staff of the school where it was picked up by the investigator.

Table V gives data showing the possible number (277) of responses which could have been received from each school and the actual number (222) of responses obtained which was slightly over eighty percent of the total possible.

TABLE V

SHOWS THE POSSIBLE RESPONSES AND RESPONSES RECEIVED FROM EACH SCHOOL STAFF TO THE TEACHER-TEACHER SOCIAL DISTANCE SCALE

School	Possible Responses	Responses Received
A	8	4
B	11	8
C	9	8
D	13	9
E	10	6
F	10	10
G	26	21
H	13	12
I	22 (59)	19
J	18	16
K	27	22
L	11	9
M	14	11
N	23	17
O	12	7
P	5	5
Q	10	5
R	45 (74)	33

The only exceptions to all of the teaching staff members receiving social distance scales for rating were in Schools I and R. It was decided that, because of the large number of persons on the teaching staffs (59 and 74 respectively), ratings should only be gathered from the teachers belonging to subject matter fields or departments in which the persons in the investigation taught. Subsequently, twenty-two teachers in School I and forty-five teachers in School R were included in the ratings. A special flyer was attached to the social distance scales in each of these schools explaining why only a limited number of teachers was used. The reasons for this were: (1) that individual teachers in so large a teaching group may not have had opportunity to become acquainted well enough with other staff members to rate them, and (2) the largeness of the two groups did not make it seem feasible that individual members would cooperate too willingly in rating so many other people even if they knew them.

The purpose in collecting the data was to obtain, as has been stated earlier, social distance ratings of teachers studied and, in particular schools, to ascertain the modified sociometric modes of the teaching group. The ratings were kept strictly anonymous with exception of knowledge of the identity of responses by the persons in the investigation. In this way it was felt that the staffs' ratings would be more valid and reliable.

Teacher-Student and Student-Teacher Social Distance Scales

These two social distance scales were adapted and modified for particular use from the Classroom Social Distance Scale mentioned earlier. The Teacher-Student and Student-Teacher Social Distance Scales are almost identical in form, with minor changes in directions and wording, to the Teacher-Teacher Social Distance Scale. Because of their similarities, they are considered together in this section.

During the first interview a copy of the Teacher-Student Social Distance Scale was given to each of the forty-one teachers together with an envelope. Upon rating each one of the classes he taught, among which was the particular class of students selected for this aspect of study and the audio-recording described earlier, the teacher delivered the envelpe sealed into the hands of the investigator. From his rating, for the students in the class which was recorded, a score was obtained which was used in the study. This score ranged in number from one to five, with one indicating the higher score.

The investigator administered the Student-Teacher Social Distance Scale to the students in the classes which had been selected for later audio-recordings. Students rated all of the teachers they were being instructed by during the second semester of the school year 1952-1953. The teachers who were rated varied due to different courses taken by students, but the one constant was the rating of the teacher being studied and in whose classroom the rating was being done. Eighty-nine percent of the possible responses were obtained, or 1018 student ratings out of a possible 1139 used in the investigation.

From the students' ratings of the teachers in the investigation, over-all scores were obtained for each and used in the study. The scores were between one and five, with one indicating the higher score.

It would be well to point out again that three kinds of scores or sets of choices were obtained from each of the three Social Distance Scales. Page two produced responses in the form of ratings of social acceptance and distance while page three of all the scales had questions which asked for responses in the form of first, second or third choices of friends and of persons who were most socially competent.

Each of the social distance scales which have been described in detail above was designed to get the same basic kinds of information from the teachers, one class of their students and from colleagues. Thus various kinds of social distance scales have the following designations so as to avoid confusion. Hereafter, the scales are referred to as S. D. S. and are preceded by the words Teacher-Teacher, Teacher-Student or Student-Teacher. If Teacher-Teacher S. D. S. is used it means that it is a school staff response, Teacher-Student S. D. S. means it is the teacher's response for his students, and Student-Teacher S. D. S. contain responses by students. Further, one instrument, the Teacher-Teacher S. D. S. was used to gather responses from both the teacher and the staff. Therefore, in order to keep the teacher's response separate from the staff responses, behind the designation of Teacher-Teacher S. D. S. is parenthetically either the word (Subject) or (Staff) standing for teacher and staff responses respectively.

The Minnesota Multiphasic Personality Inventory

The Minnesota Multiphasic Personality Inventory, hereafter referred to as the M. M. P.I., was chosen as the personality measuring instrument from which certain measures of personally rated social adjustment could be abstracted and used in this study. The M. M. P. I. was chosen because of the following reasons: (1) a survey of the existing personality instruments was made and experts' critical analyses, such as contained in Buros (16), of them were read; the literature tended to reveal that the M. M. P. I. was as good or better than other such inventories and was more valid and reliable, (2) the comparative ease and facility with which administration of the instrument could be accomplished with forty-one teachers in eighteen different locations, (3) the test lent itself to the combining of more than one of its scales into a composite termed "social competence," and (4) the M. M. P. I. yields, among other trait scores, a single quantitative score on an additional trait known as Social Introversion (SI) which was also used in this investigation.

The prevalent opinion is that the subjective, structured, paper and pencil personality inventories are at best not to be relied upon to any great extent. The M. M. P. I. however, is recognized as being among the better available instruments. Ellis (21) states:

> Among the newer questionnaires the M. M.

> P. I. appears to be the most promising, perhaps because it gets away from the type of administration which has hitherto been almost synonymous with personality testing.

Richards (37) corroborates this view in one of the more recent reviews of the validation of personality inventories as follows:

> The Minnesota Scale is in our opinion the best of the exhaustive questionnaire type scales; it is that most readily used by the patient and most easily scored; it permits a focus clinically on those answers that are of most interest, i. e., the atypical answers.

Ellis (21) further reports that out of fifteen validations of the M. M. P. I. ten were positive validations, three were questionable, and two were negative. Of four other personality inventories compared with the M. M. P. I. the nearest inventory to approach it was the Woodworth Personal Data Sheet with only eleven positive validations out of twenty-nine, four being questionable and fourteen negative.

The M. M. P. I. gives scores on four validating scales and the nine basic personality scales. The four validating scales are a question scale, lie scale, validity or (F) scale, and a test attitude of (K) scale. The personality scales on which scores can be calculated are those for Hypochondriasis, Depression, Hysteria, Psychopathic Deviate, Masculinity-Femininity of Interests, Paranoia, Psychasthenia, Hypomania, and Schizophrenia.

The scales are based upon clinical cases classified according to conventional psychiatric nomenclature. The descriptions follow.

The Question Score (?): The Question score is a validating score consisting simply of the total number of items put in the "Cannot say" category. The size of this score affects other scores.

The Lie Score (L): The Lie score is also a validating score that affords a measure of the degree to which the subject may be attempting to falsify his scores by always choosing the response which places him in the most acceptable light.

The Validity Score (F): The Validity score is not a personality scale, but serves as a check on the validity of the whole record. If the F score is high, the other scales are likely to be invalid either because the subject was careless, or unable to comprehend the items, or because someone made extensive errors in entering the items on the record sheet.

The K Score (K): The K score is used essentially as a correction factor to sharpen the discriminatory power of the clinical variables measured by the Inventory. As such, K acts as a suppressor variable.

The Hypochrondriasis Scale (Hs): The Hs scale is a measure of amount of abnormal concern over bodily functions.

The Depression Scale (D): The D scale measures the extent of the clinically recognized symptom complex, depression.

The Hysteria Scale (Hy): The Hy scale measures the degree to which the subject is like patients who have developed conversion-type hysteria symptoms.

The Psychopathic Deviate Scale (Pd): The Pd scale measures the similarity of the subject to a group of persons whose main difficulty lies in their absence of deep emotional response, their inability to profit from experience, and their disregard for social mores.

The Interest Scale (Mf): This scale measures the tendency toward masculinity or femininity of interest pattern; a high score indicates an interest pattern corresponding to that of the opposite sex.

The Paranoia Scale (Pa): The Pa scale was derived by contrasting normal persons with a group of clinic patients who were characterized by suspiciousness, oversensitivity, and delusions of persecution.

The Psychasthenia Scale (Pt): The Pt scale measures the similarity of the subject to psychiatric patients who are troubled by phobias and compulsive behavior.

The Schizophrenia Scale (Sc): The Sc scale measures the similarity of the subject's responses to those patients who are characterized by bizarre and unusual thoughts or behavior.

The Hypomania Scale (Ma): The Ma scale measures the personality factor characteristic of persons with marked overactivity in thought and action.

The basic concept of the Inventory assumes that among the entire 550 items there are groupings of items that can form additional scales. The K-factor scale was developed in such a manner. An additional scale to those above, and one from which quantitative scores are used in this investigation, is the Social I. E. Scale (Si) which aims to measure the tendency to withdraw from social contact with others. This scale is especially valuable for use with normals; under such heading the teachers in this study are assumed to belong.

The M. M. P. I. is copyrighted in an individual form and in a pencil and paper group form. Decision was made to use the group form and, subsequently, each of the forty-one teachers received a copy of the test in group form together with an answer sheet designed for machine scoring. Since the Inventory is not timed, and since the directions appear on the front cover of the test booklet, no instructions were offered the subjects. Test booklets and completed answer sheets were collected during the next visitation. The responses were machine scored and individual profiles made of each teacher by the Student Counseling Center at the University of Wisconsin. The data and interpretations placed on them are confidential so they do not directly appear in the investigation. M. M. P. I. profiles yield estimates directly in T-scores for the Social Introversion (Si) scale which were used in the investigation. The lower the score the less there is tendency to social introversion or withdrawal from social contact with others.

After consultation with two qualified clinicians experienced in the use of the M. M. P. I. as an instrument for personality diagnosis, it was mutually decided that another measure, which was termed "Social Competence," could be obtained from a composite interpretation of related traits as shown on the teachers' profiles. "Social Competence" here was considered to mean the same s defined in Section I, page 100. The rank ordering of teachers, using the interpretations of the two qualified clinicians, for "social competence" was based on the following assumptions: (1) that the records belong to a so-called "normal" group of individuals, who at the time had no gross physical or mental illness, (2) that the more free the individual is from neurotic conflicts, the better adjusted he is, and therefore the more competent in his social environmental interactions, and (3) that excessive "socializing" may be a function of anxiety, or in other words, an individual may "socialize" without being "competent."

The records were first dichotomized, in the procedure for placing the teachers in rank order, according to their deviations from the mean T-score. Those who had scores of two or more sigmas above or below the mean were placed in a lower group, on the basis of the statistics given by one of the authors of the instrument that T-scores above seventy occur in only one out of seven records of normal people.

The extent and nature of the deviating columns were taken into consideration next. Due to the particular nature of the Mf and Pa columns, a great deal of caution was used in their interpretation. Interpretations were based on the configurations rather than single elevations in any single column. According to statistics given by Meehletal in his "Pattern Analysis of the Minnesota Multiphasic Personality Inventory", the most usual peaks in normal persons are on: (1) the Ma (Hypomania Scale) plus the Pd (Psychopathic Deviate Scale); (2) Ma (Hypomania Scale); (3) Hy (Hysteria Scale); and (4) Pt (Psychasthenia Scale) plus Pd (Psychopathic Deviate Scale). The profiles were judged accordingly on the basis of the above scales peaks, always taking into consideration the extent of the individual peak and its relationship to the other relatively high and low points from the view of "social competence."

The records were judged finally for the Si (Social Introversion Scale) scores and the extent to which the "socialization," as indicated by this column, would be successful and normally gratifying to the subject and those around him, on the basis of the personality make-up obtained from the configurations in the aforementioned columns. Thus there was a good deal of re-shuffling after the initial grouping; the final ranking is the result of a composite of all the named factors which could indicate "social competence." The rank order of the forty-one teachers was converted to individual T-scores which were used in the investigation. The higher the T-score the more socially competent the individual is.

Audio-Recorded Teaching Presentation Social Ratings

Earlier in the investigation it was stated that an audio-recording was taken of a selected teaching presentation. It has also been previously explained that subsequent to the recording, a rating scale was constructed which was divided into two different sections. One section was comprised of items purporting to measure professional teaching competence while the other section was composed of items measuring social competence. The same recordings of teaching presentations which were rated for teaching competence by the two trained persons were also simultaneously rated for social competence. The method of attaining the single score for each teacher was the same as stated in the section devoted to Audio-Recorded Teaching Presentation Professional Ratings, page 105. The low score indicates the high rating.

Reliability and Validity of Social Competence Criteria

A reliability estimate was obtained between the two raters who estimated the social competence of the forty-one subjects from the audio-recordings. The data reveals that there exists a high relationship (+.99) between the raters using the same instrument to measure the same recordings. No other reliability estimates were made of the other instruments measuring social competence and interaction. Table VI contains the intercorrelations of the five instruments used to measure social factors.

The relationships shown in Table VI which exist between measures of social factors are not too high but it would seem reasonable to assume that the opinions of the raters were reliable and valid at the time of rating and for the instruments used.

To summarize briefly, the discussion in this section has been devoted to a description of the design of the investigation, how the selection of the forty-one teachers was made, the six criteria of teaching success and seven measures of social factors, and the method by which the data were collected. Attempt is made where possible to show the validity and reliability of the instruments and the representativeness of the selected population.

SECTION III

ANALYSIS OF THE DATA

THE COLATION and analysis of these data were accomplished according to procedures set forth in the materials to follow. After all of the raw scores for the various criteria for teaching success and social competence had been assembled and placed in an over-all table, it was decided that the State Department X-Ratings, Administrator M-Blank Ratings, and the investigator's M-Blank Rating would be combined into a single composite score. This was accomplished for each teacher in the study by first totaling and then averaging the two M-Blank scores. The resultant average M-Blank score was then totaled and averaged with the State Department X-Rating Score. A final composite score was derived.

Intercorrelations were then calculated and coefficients obtained between each of the four teaching efficiency measures and the seven social competency measures for a total of fifty-five correlations. Significant relationships of +.300 or above were found in twenty-five of the fifty-five correlations, and twelve were +.500 or higher.

It was next ascertained if mean differences existed between the most successful teachers and least successful teachers. It was arbitrarily decided that approximately one-third, or fourteen persons, would comprise an upper group and fourteen people a lower group. The four teaching efficiency criteria were employed consecutively to select a different group each of 14 high and 14 low teachers.

Comparisons were made between pairs of teachers, comparing the first most successful and least successful teacher followed by the next most successful teacher and next least successful teacher, etc. The most successful teachers were higher a total of 227 times, even fifty-five times and lower 110 times than the least successful teachers in paired mean differences on all seven measures of social competence. Where the teacher self-evaluation was used to select the fourteen upper and fourteen lower teachers, the fourteen least successful teachers here provide the only exception of achieving higher paired mean differences.

Modified sociograms of friendship and social competence choices were tabulated from answers to the last two questions of page 3 of the social distance scales. After all of the first, second and third choices were totaled for both social competence and friendship, it was decided that modified sociograms would be drawn for illustration purposes only of the three most successful teachers and the three least successful teachers with their respective school staffs. Identification of the three most and three least successful teachers was made from the four tables of mean differences. The chosen teachers were those who appeared most consistently, without reversal to the opposite group, in their respective high or low groups. Once selected, sociograms were drawn of these teachers in relation to their respective school staffs and are shown on the next several pages with accompanying explanations.

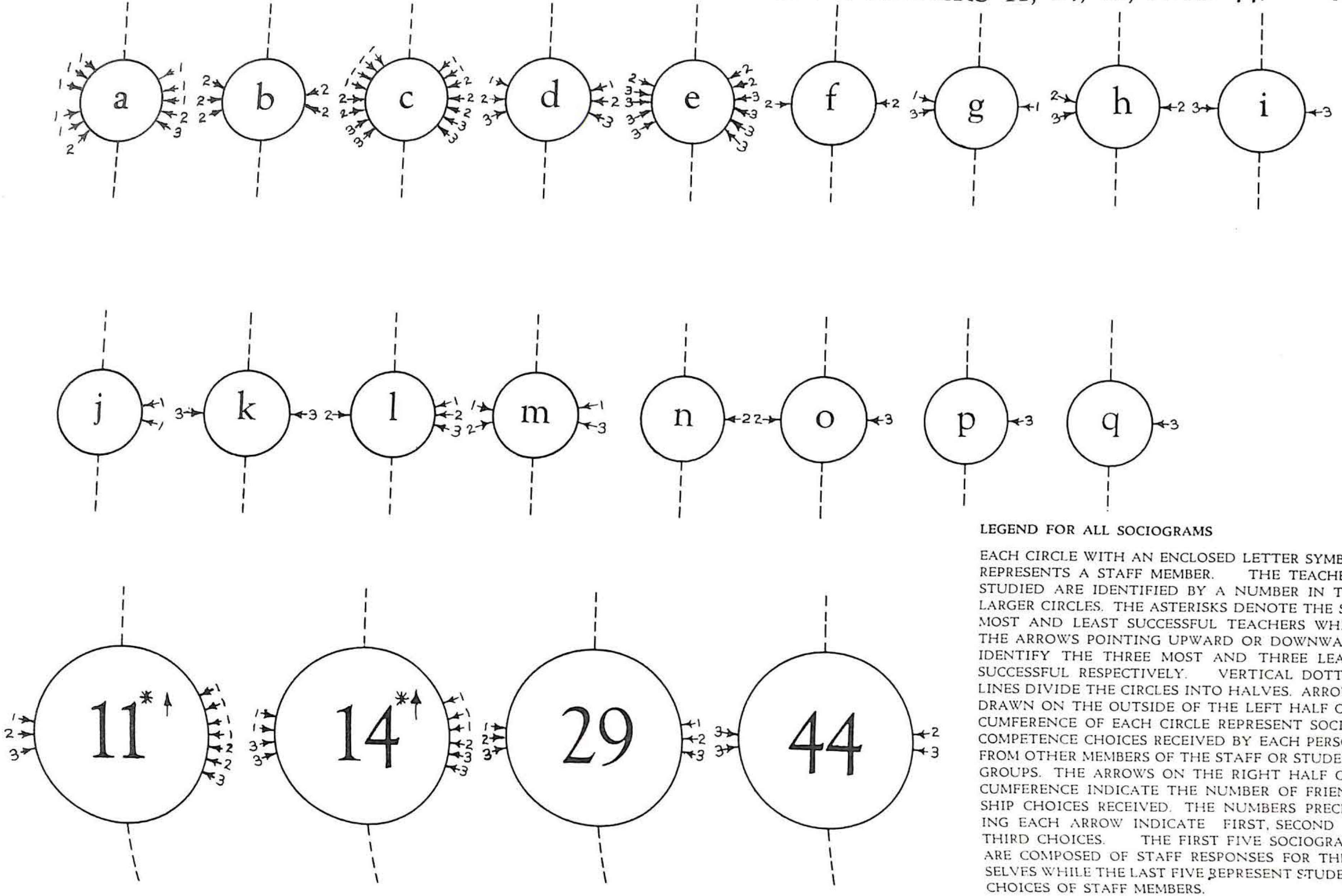
The modified sociograms show that teachers 11, 14 and 20 received a total of sixty-six first, second and third choices in social competence from their students out of a possible total of 354 responses whereas teachers 2, 45 and 47 received a total of ten choices out of a possible 171. The most successful teachers, 11, 14, and 20 received twelve first, second and third choices in social competence out of a possible total of 138 school staff responses. The three least successful teachers, 2, 45 and 47 were chosen two times by their school staffs whose first, second and third choices in social competence totaled 117.

The three most successful teachers received totals of seventeen school staff and fifty-four student friendship choices out of 138 and 354 total possible responses respectively. Three school staff and ten student friendship choices were received by the three least successful teachers from possible total responses of 117 and 171 respectively.

When compared with individual members of their respective teaching staffs in social competence and friendship choices, the three most successful teachers appear to be among the leaders or at least have sub-leader status. The three least successful teachers, on the other hand are relative fringers and isolates.

Tables VII, VIII, IX and X present analyses in percentage of the number of friendship and social competence responses each of the six teachers received and a percentage comparison showing differences between the three most and three least successful teachers. The three most successful teachers, 11, 14 and 20, received 10.4% and 12.9% whereas the three least successful teachers, 2, 45 and 47 obtained .98% and 7.4% of the total staff and student social competence

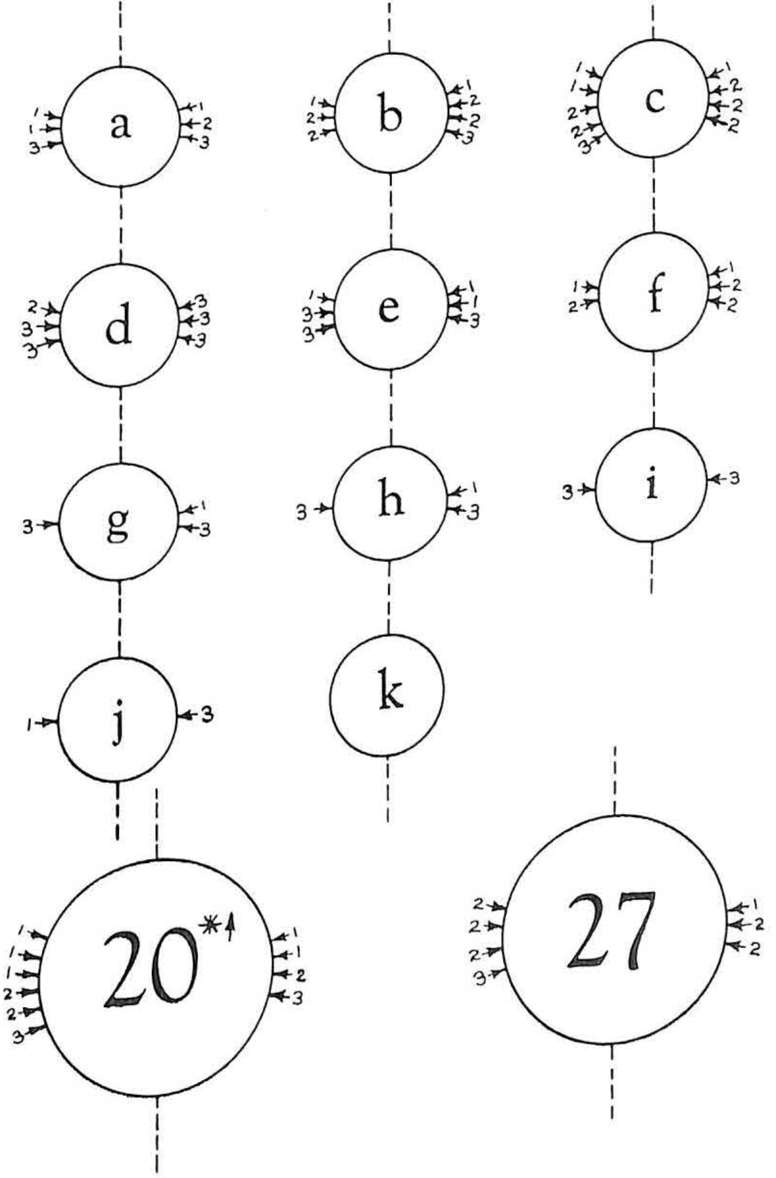
MODIFIED SOCIOGRAM OF THE STAFF IN SCHOOL I DEPICTING PARTIAL STAFF SOCIAL COMPETENCE AND FRIENDSHIP CHOICES FOR TEACHERS 11, 14, 29, AND 44.



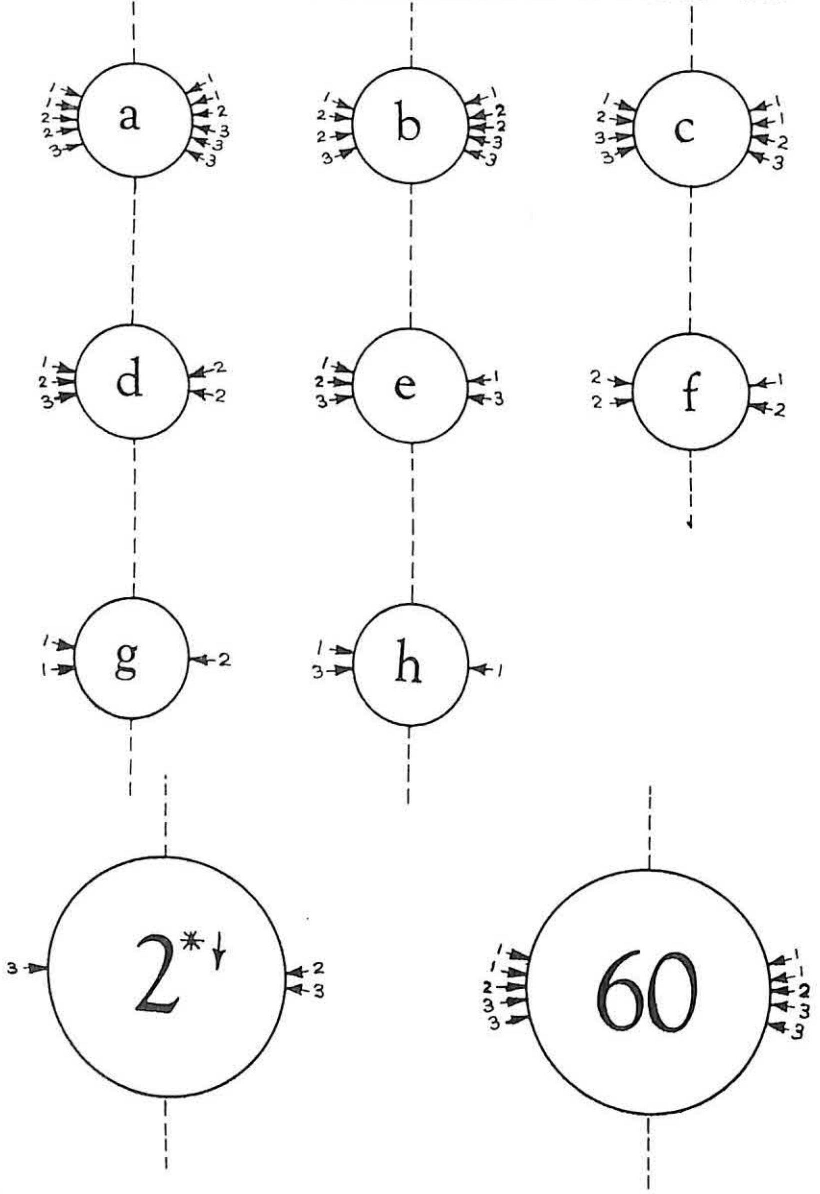
LEGEND FOR ALL SOCIOGRAMS

EACH CIRCLE WITH AN ENCLOSED LETTER SYMBOL REPRESENTS A STAFF MEMBER. THE TEACHERS STUDIED ARE IDENTIFIED BY A NUMBER IN THE LARGER CIRCLES. THE ASTERISKS DENOTE THE SIX MOST AND LEAST SUCCESSFUL TEACHERS WHILE THE ARROWS POINTING UPWARD OR DOWNWARD IDENTIFY THE THREE MOST AND THREE LEAST SUCCESSFUL RESPECTIVELY. VERTICAL DOTTED LINES DIVIDE THE CIRCLES INTO HALVES. ARROWS DRAWN ON THE OUTSIDE OF THE LEFT HALF CIRCUMFERENCE OF EACH CIRCLE REPRESENT SOCIAL COMPETENCE CHOICES RECEIVED BY EACH PERSON FROM OTHER MEMBERS OF THE STAFF OR STUDENT GROUPS. THE ARROWS ON THE RIGHT HALF CIRCUMFERENCE INDICATE THE NUMBER OF FRIENDSHIP CHOICES RECEIVED. THE NUMBERS PRECEDING EACH ARROW INDICATE FIRST, SECOND OR THIRD CHOICES. THE FIRST FIVE SOCIOGRAMS ARE COMPOSED OF STAFF RESPONSES FOR THEMSELVES WHILE THE LAST FIVE REPRESENT STUDENT CHOICES OF STAFF MEMBERS.

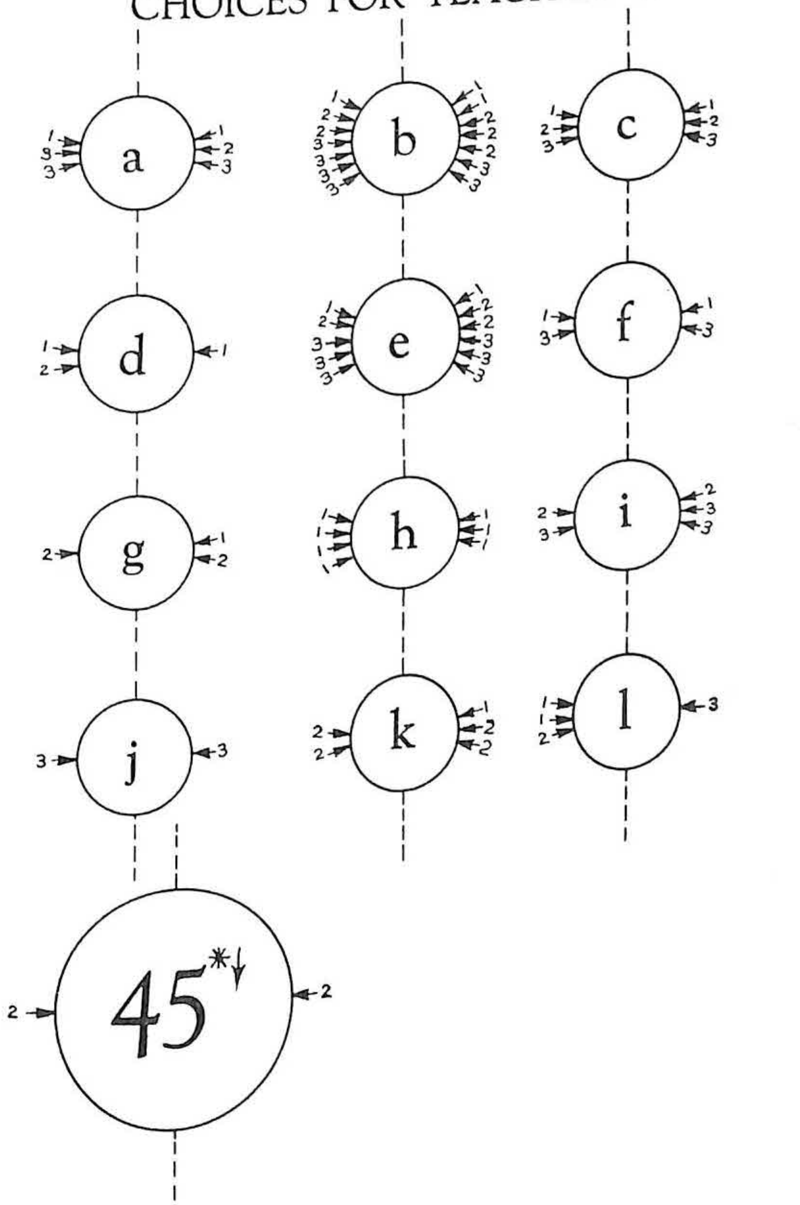
MODIFIED SOCIOGRAM OF THE STAFF IN SCHOOL M DEPICTING PARTIAL STAFF SOCIAL COMPETENCE AND FRIENDSHIP CHOICES FOR TEACHERS 20 AND 27.



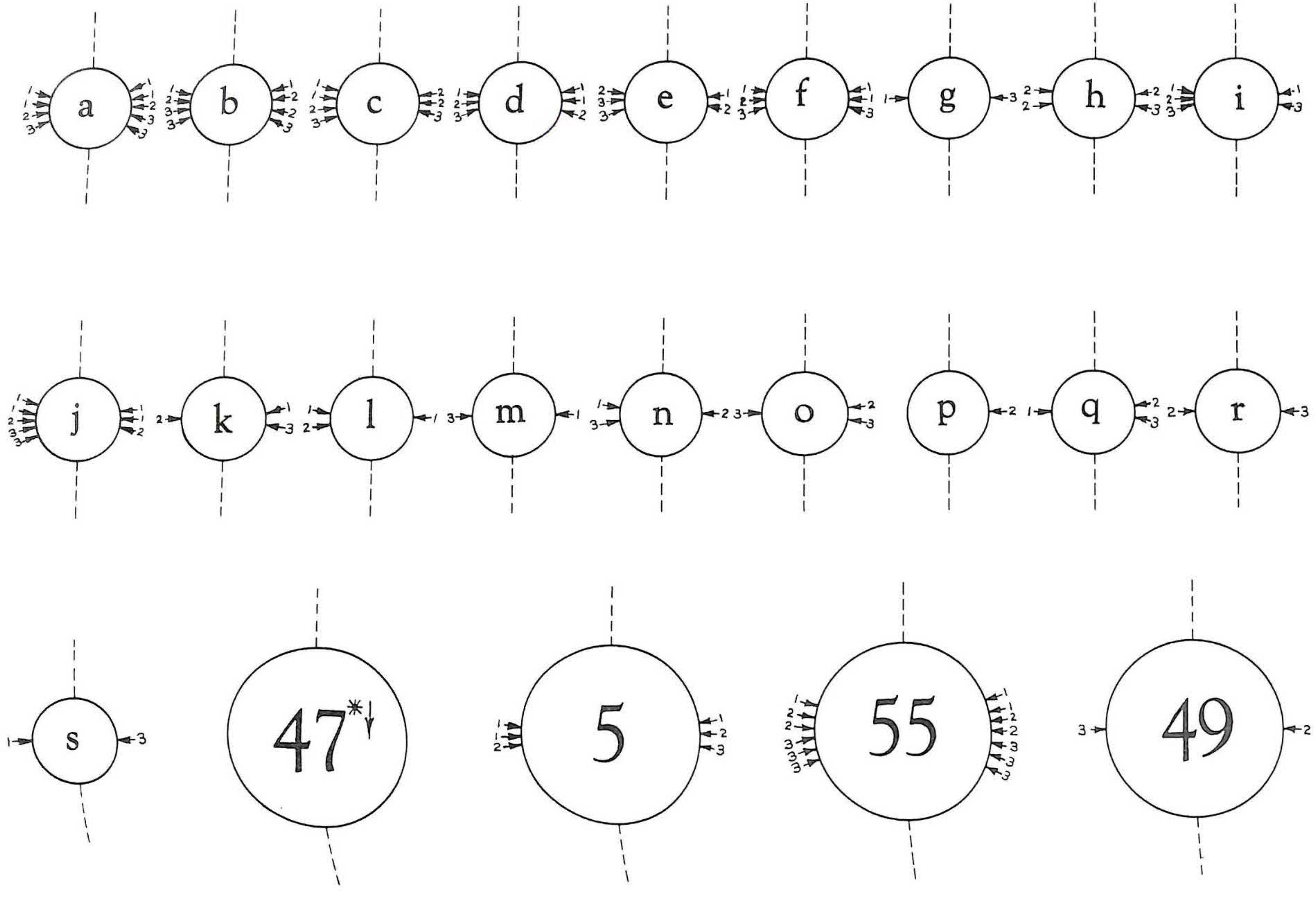
MODIFIED SOCIOGRAM OF THE STAFF IN SCHOOL F DEPICTING PARTIAL STAFF SOCIAL COMPETENCE AND FRIENDSHIP CHOICES FOR TEACHERS 2 AND 60.



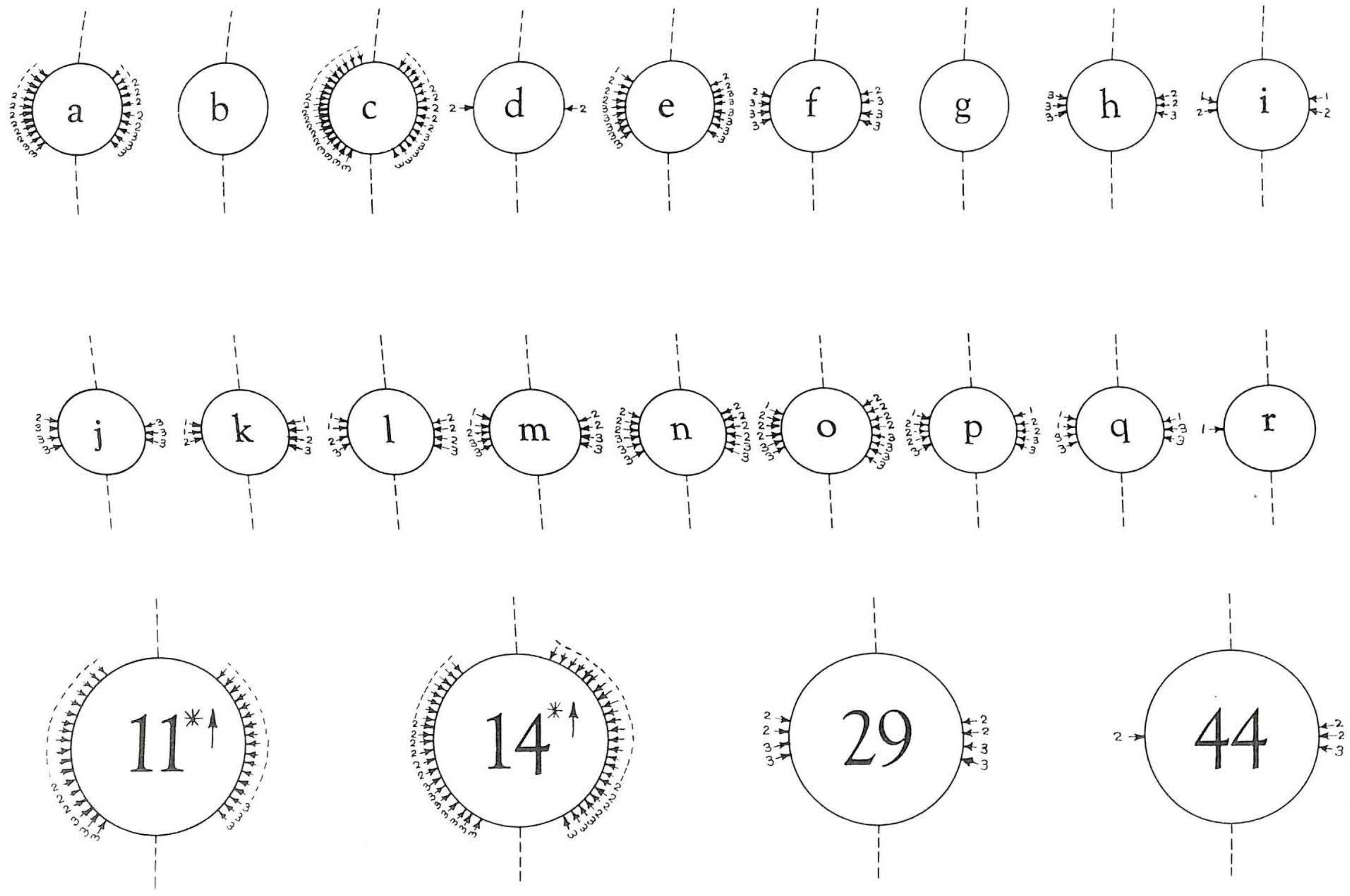
MODIFIED SOCIOGRAM OF THE STAFF IN SCHOOL H DEPICTING PARTIAL STAFF SOCIAL COMPETENCE AND FRIENDSHIP CHOICES FOR TEACHER 45.



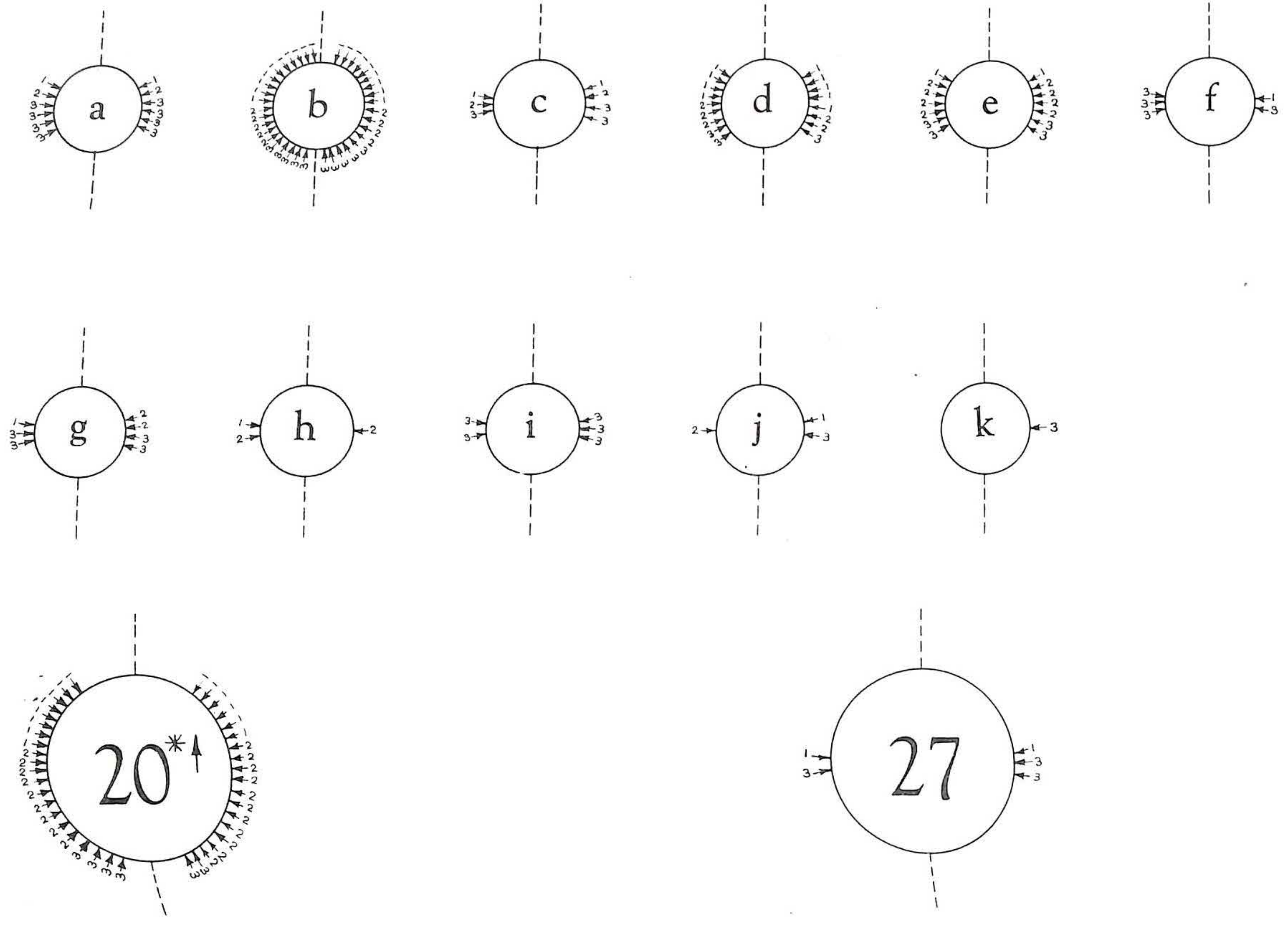
MODIFIED SOCIOGRAM OF THE STAFF IN SCHOOL N DEPICTING PARTIAL STAFF SOCIAL COMPETENCE AND FRIENDSHIP CHOICES FOR TEACHERS 47, 5, 55, AND 49.



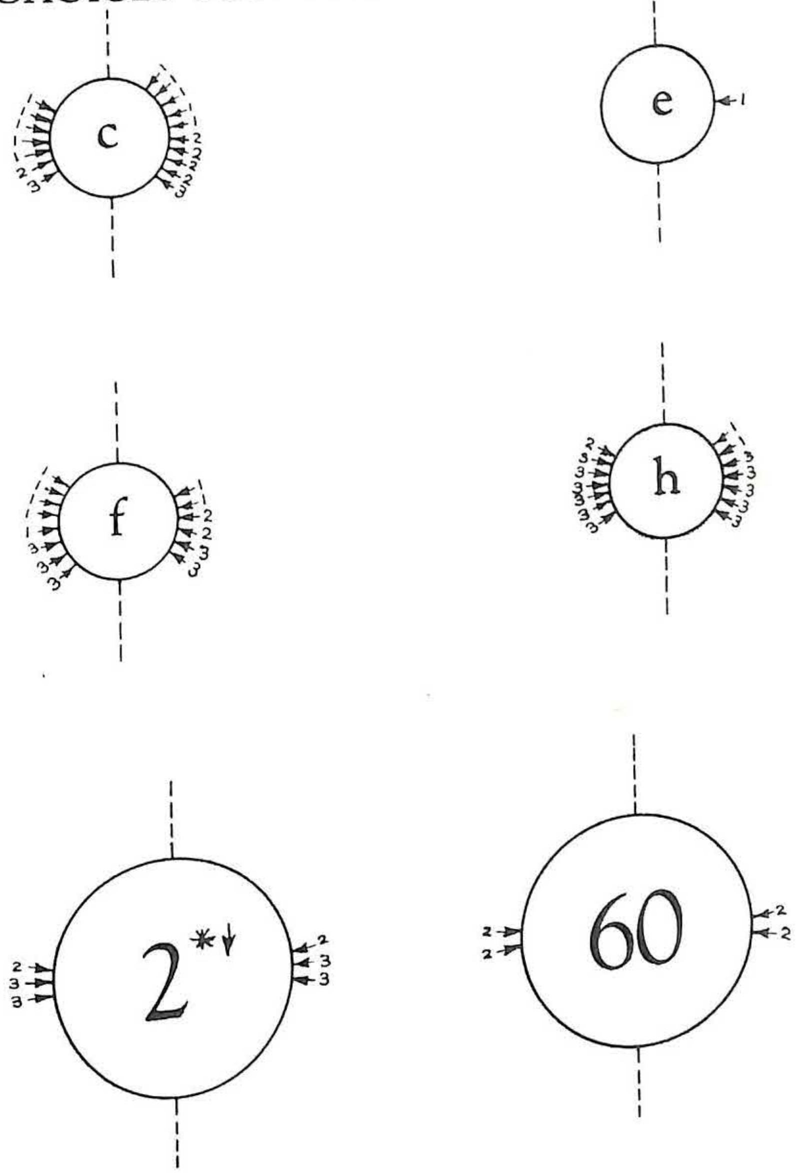
MODIFIED SOCIOGRAM OF PARTIAL SCHOOL STAFF IN SCHOOL I DEPICTING STUDENT SOCIAL COMPETENCE AND FRIENDSHIP CHOICES FOR TEACHERS 11, 14, 29, AND 44.



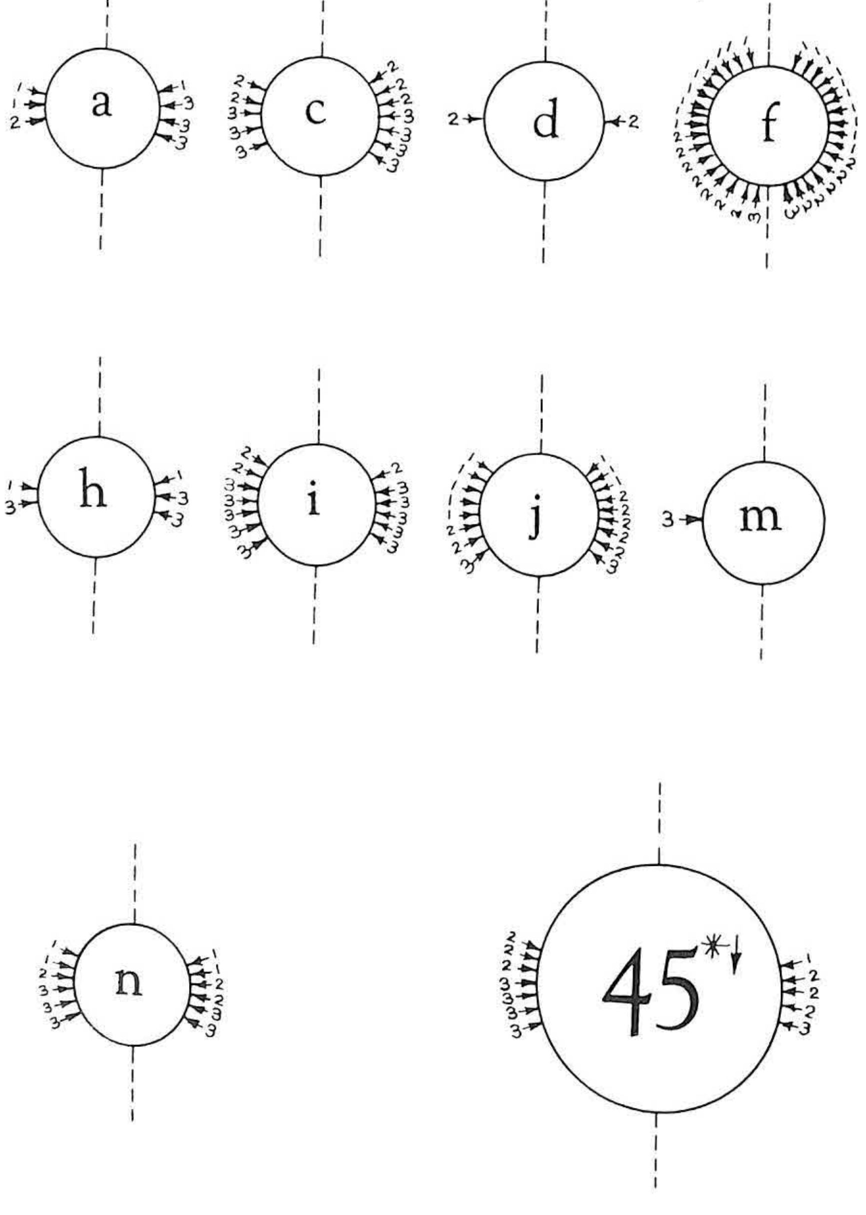
MODIFIED SOCIOGRAM OF PARTIAL SCHOOL STAFF IN SCHOOL M DEPICTING STUDENT SOCIAL COMPETENCE AND FRIENDSHIP CHOICES FOR TEACHERS 20 AND 27.



MODIFIED SOCIOGRAM OF PARTIAL SCHOOL STAFF IN SCHOOL M DEPICTING STUDENT SOCIAL COMPETENCE AND FRIENDSHIP CHOICES FOR TEACHERS 2 AND 60



MODIFIED SOCIOGRAM OF PARTIAL SCHOOL STAFF IN SCHOOL H DEPICTING STUDENT SOCIAL COMPETENCE AND FRIENDSHIP CHOICES FOR TEACHER 45.



MODIFIED SOCIOGRAM OF PARTIAL SCHOOL STAFF IN SCHOOL N DEPICTING STUDENT SOCIAL COMPETENCE AND FRIENDSHIP CHOICES FOR TEACHERS 47, 5, 55, AND 49.

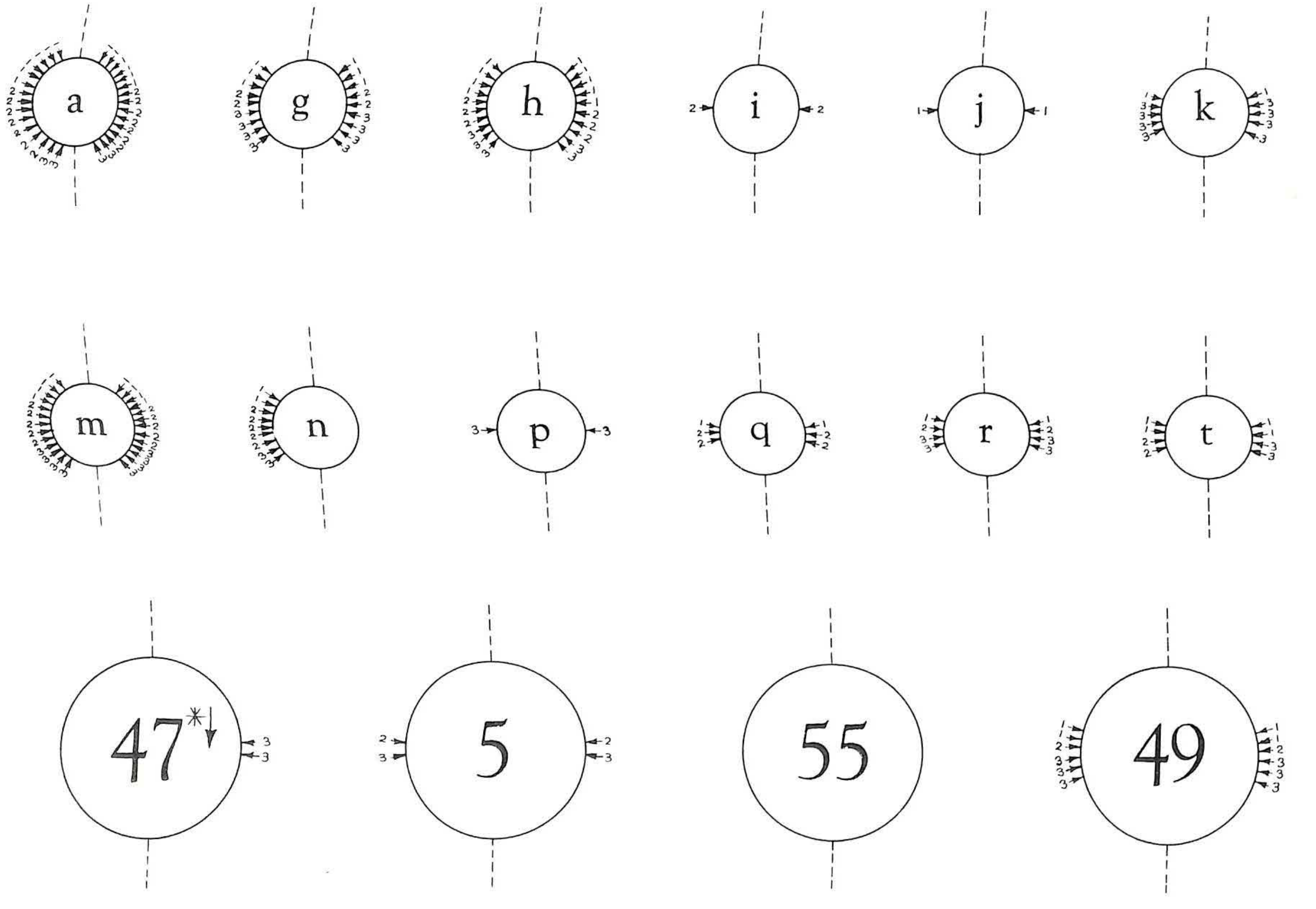


TABLE VI

INTERCORRELATIONS OF SOCIAL CRITERIA RATINGS

Rating	Correlation 1	2	3	4	5
Teacher-Teacher Social Distance Scale	.....	.....	.....	.....	.....
Teacher-Student Social Distance Scale	+.345	.....	.....	.....	.....
Student-Teacher Social Distance Scale	+.591	+.193	.....	.....	.....
Recording Ratings of Social Competence	+.278	+.315	+.354	.....	.....
M. M. P. I. (Social Competence Scale)	-.385	-.282	-.477	-.503	.....
M. M. P. I. (Social Introversion Scale)	-.186	-.077	-.104	-.007	-.093

Note: The coefficients for the M. M. P. I. (Social Competence Scale) with minus signs are actually positive correlations.

TABLE VII

DEPICTING DATA WHICH SHOWS THREE MOST SUCCESSFUL AND THREE LEAST SUCCESSFUL TEACHERS AND STAFF SOCIAL COMPETENCE CHOICES RECEIVED BY THEM

School	Teacher No.	Social Competence 1st Choices No. Rec'd	Possible No.	No. in %	Social Competence 2nd Choices No. Rec'd	Possible No.	No. in %	Social Competence 3rd Choices No. Rec'd	Possible No.	No. in %
				Three Most Successful Teachers						
I	11	1	19	5%	1	19	5%	0	19	0%
I	14	2	19	10%	0	19	0%	2	19	10%
M	20	3	11	27%	2	11	18%	1	11	9%
Average Total				14%			9.3%			6.3%
				Three Least Successful Teachers						
F	2	0	10	0%	0	10	0%	1	10	1%
H	45	0	12	0%	1	12	8%	0	12	0%
N	47	0	17	0%	0	17	0%	0	17	0%
Average Total				0%			2.6%			.33%
Average Grand Total of Staff Social Competence Choices received by the Three Most Successful Teachers										10.4%
Average Grand Total of Staff Social Competence Choices Received by the Three Least Successful Teachers										.98%

TABLE VIII

DEPICTING DATA WHICH SHOWS THREE MOST SUCCESSFUL AND THREE LEAST SUCCESSFUL TEACHERS AND STUDENT SOCIAL COMPETENCE CHOICES RECEIVED BY THEM

		Social Competence 1st Choices			Social Competence 2nd Choices			Social Competence 3rd Choices		
School	Teacher No.	No. Rec'd	Possible No.	No. in %	No. Rec'd	Possible No.	No. in %	No. Rec'd	Possible No.	No. in %
				Three Most Successful Teachers						
I	11	12	43	28%	3	43	7%	3	43	7%
I	14	8	43	18%	6	43	14%	8	43	18%
M	20	10	32	3%	10	32	3%	6	32	19%
Average Total				16%			8%			14.6%
				Three Least Successful Teachers						
F	2	0	10	0%	1	10	10%	2	10	20%
H	45	0	19	0%	3	19	16%	4	19	21%
N	47	0	28	0%	0	28	0%	0	28	0%
Average Total				0%			8.6%			13.6%
Average Grand Total of Student Social Competence Choices Received by the Three Most Successful Teachers										12.9%
Average Grand Total of Student Social Competence Choices Received by the Three Least Successful Teachers										7.4%

TABLE IX

DEPICTING DATA WHICH SHOWS THREE MOST SUCCESSFUL AND THREE LEAST SUCCESSFUL TEACHERS AND SCHOOL STAFF FRIENDSHIP CHOICES RECEIVED BY THEM

		Friendship 1st Choices			Friendship 2nd Choices			Friendship 3rd Choices		
School	Teacher No.	No. Rec'd	Possible No.	No. in %	No. Rec'd	Possible No.	No. in %	No. Rec'd	Possible No.	No. in %
				Three Most Successful Teachers						
I	11	4	19	21%	2	19	10%	1	19	5%
I	14	3	19	16%	1	19	5%	2	19	10%
M	20	2	11	18%	1	11	9%	1	11	9%
Average Total				18%			8%			8%
				Three Least Successful Teachers						
F	2	0	10	0%	1	10	10%	1	10	10%
H	45	0	12	0%	1	12	8%	0	12	0%
N	47	0	17	0%	0	17	0%	0	17	0%
Average Total				0%			6%			3.3%
Average Grand Total of Staff Friendship Choices Received by Three Most Successful Teachers										11.3%
Average Grand Total of Staff Friendship Choices Received by Three Least Successful Teachers										3.1%

TABLE X

DEPICTING DATA WHICH SHOWS THREE MOST SUCCESSFUL AND THREE LEAST SUCCESSFUL TEACHERS AND STUDENT FRIENDSHIP CHOICES RECEIVED BY THEM

		Friendship 1st Choices			Friendship 2nd Choices			Friendship 3rd Choices		
School	Teacher No.	No. Rec'd	Possible No.	No. in %	No. Rec'd	Possible No.	No. in %	No. Rec'd	Possible No.	No. in %
Three Most Successful Teachers										
I	11	14	43	32%	0	43	0%	3	43	7%
I	14	17	43	39%	4	43	9%	3	43	7%
M	20	8	32	24%	2	32	6%	3	32	9%
Average Total				31. 7%			5%			7. 6%
Three Least Successful Teachers										
F	2	0	10	0%	1	10	10%	2	10	20%
H	45	11	19	5%	3	19	16%	1	19	5%
N	47	0	28	0%	0	28	0%	2	28	7%
Average Total				1. 7%			8. 6%			10. 6%
Average Grand Total of Student Friendship Choices Received by Three Most Successful Teachers										14. 7%
Average Grand Total of Student Friendship Choices Received by Three Least Successful Teachers										7 %

choices respectively. For friendship responses from the staffs and students, the upper three teachers received totals of 11.3% and 14.7% while the lower teachers obtained totals of 3.1% and 7.0% respectively.

SECTION IV

SUMMARY OF FINDINGS AND CONCLUSIONS

THE PERSONS studied in this investigation were forty-one in-service teachers in eighteen Wisconsin Senior High Schools, and were graduates from the School of Education of the University of Wisconsin between the years 1940-1952 inclusive. All of the teachers were located in public educational institutions within a defined geographical area bounded by a circle whose radius extended fifty miles from Madison, Wisconsin.

The normative-survey method of investigation was followed in this study. The instruments employed yielded data which fell into two major categories: data that were relative to teaching success and data that were relative to social competence, social interaction and group structure.

Success in teaching was identified by obtaining an over-all judgment for each individual by using the following composite teaching criteria: (1) teacher self-ratings, (2) adapted M-blank ratings by local administration, (3) state department supervisory X-ratings, (4) student's ratings of teachers, (5) opinion of a trained observer, and (6) ratings by qualified persons of classroom teaching through audio-recordings. Data regarding social aspects were identified by: (1) sociometric techniques modified for use in classrooms and with staff and students, (2) opinions of judges of social competency as heard in audio-recordings of teaching presentations, and (3) teacher self-rated social adjustment and competency tests.

Inter-correlations were obtained for all of the variables measured. Mean differences of the fourteen most successful and fourteen least successful teachers were calculated. The three most and least successful teachers were next identified and modified sociograms showing their friendship and social competence status in relation to their respective teaching staffs were constructed.

The purpose of the present investigation as initially stated in section one was to determine the relationship between success in teaching and certain factors of social interaction, individual social competence and inter-personal-group modes. Certain questions seemed related to the problem. These together with the conclusions reached are as follows:

1. Does the concurrent use of instruments such as sociometric techniques, audio-recording ratings, observation, personal adjustment inventories, and teacher ratings provide worthwhile measures of teaching success?

Summary

Coefficients of correlation were obtained for all of the various instruments used in this study. Twenty-five of the fifty-five correlations show significant relationships (+.300 or above) between each other and twelve coefficients are +.500 or more. Social distance scores have been ascertained for each teacher and for each teaching staff to which he belongs. Twenty-nine of the forty-one teachers in the investigation have indicated by their social distance scores, which are lower than the teaching staffs' scores of them, that they accept their colleagues more than they are themselves accepted by the groups. Equal or mutual acceptance by both the staff and the teacher is found in two cases, teachers 48 and 53. The other ten teachers accept their colleagues less than the staffs accept them. Modified sociograms of the teaching staffs and one class each of the three most successful and three least successful teachers are to be found on pages 112 through 121. Patterns of group structure were found, and individual group leaders, fringers, and isolates were identified. The summary under question 2 which follows gives a more thorough explanation.

Conclusion

The concurrent use of sociometric techniques, audio-recording ratings, observational and administrative evaluations, and personal adjustment self-ratings would appear to provide partial measures of teaching success. Some of the coefficients of correlation obtained in this study in the area of social competence were such as to merit further study. The modified sociograms and social distance scores appear to provide definite insights into why the individual person may succeed or fail as a teacher. The three most successful teachers in this investigation are all, for example, in positions of comparative leadership as shown in the modified sociograms and are greatly accepted as members of the teaching group to which they belong. The three least successful teachers in the study as depicted in the sociograms are fringers or relative isolates in their respective teaching staffs and are not too well accepted socially.

2. Can definite patterns of group social structures be established for the teacher and his classes, and for the teacher and the partic-

ular school staff to which he belongs?

Summary

Examples of such group social structure patterns are shown in the modified sociograms in section III. The sociograms were specifically drawn to show the three most successful teachers (11, 14 and 20 in schools I and M), and the three least successful teachers (2, 45, and 47 in schools F, H and N), in relation to the particular class studied and staff group structures of which they were part. Teachers 29, 44, 27, 5, 55, 49, and 60 are also included in the modified sociograms because they were members of the above school staffs. Teachers 11, 14 and 20, the three most successful, received social competence choices from their students totaling sixty-six and from their teaching staffs totaling twelve out of 354 and 138 possible choices, respectively. The three most successful teachers received friendship choices from staffs and students totaling seventeen and fifty-four, whereas the three least successful teachers received three and ten total friendship choices, respectively.

The three most successful teachers appear to be teaching staff group leaders or have at least sub-group leader status, while the three least successful teachers are found to be isolates and fringers.

Conclusion

It would appear that definite patterns of group social structures can be established for the teacher and his classes as well as for the teacher and the particular teaching staff of which he is a part.

3. Is there a relationship between these group social structures and success in teaching?

Summary

Tables IX and X, Section III, contain data showing the three most successful with the three least successful teachers and the first, second, and third friendship choices received by them from colleagues and one class of their students. The average grand total of staff friendship choices received by the three most successful teachers is 11.3% whereas the three least successful teachers received only 3.1%. The average grand total of student friendship choices received by the three most successful teachers is 14.7% and that for the three least successful teachers is 7%.

Conclusion

The most successful teachers appear to be more popular and held in greater esteem as friends by their colleagues and students than are the least successful teachers. The more effective teachers occupy positions of leadership and higher status whereas the poorer teachers are isolates or have only one mutual friend on the staff.

4. Is there a relationship between the teacher's self-concept of social adjustment and his teaching success?

A. The Teacher-Teacher Social Distance Scale (Self-Rating)

Summary

Little or no relationship was found between the teacher's self-concept or social adjustment and teaching success with one exception. A low but significant correlation (+.321) exists between the teacher's self-evaluation of his teaching and his self-rating of how he believes he is socially accepted by the group of teachers of which he is a member. Slight negative relationship (-.045) is found between the composite score of Supervisory Ratings and the Teacher-Teacher Social Distance Scale while a correlation of +.133 is shown between the Pupil Reaction to Instruction responses and the self-rated Teacher-Teacher Social Distance Scale. Almost no relationship (+.032) is indicated between the Professional Recording ratings and the Teacher-Teacher Social Distance Scale (self-rated).

Conclusion

From the data presented it can be concluded that the teacher's concept of how well he feels he is accepted by his colleagues bears little or no resemblance to the degree of success he attains in teaching. The fact that there is a low but significant relationship between the teacher's own ratings of his teaching and social acceptance leads the investigator to wonder to what extent teachers are subjectively biased in all phases of their living and whether self-evaluations are critical enough to be considered valuable inclusions for measuring efficiency in teaching.

B. The Teacher-Student Social Distance Scale (Teacher Self-Rating)

Summary

Correlations (+.150, -.068, and +.166) between the Teacher-Student Social Distance Scale and the composite criteria of success in teaching are insignificant in all cases except for the Professional Recording ratings (+.441).

Conclusion

It is concluded that the teacher's estimation of how well he is accepted by his students has little to do with how well he is teaching. It is possible that in rating himself for acceptability by students he places greater weight on the oral factor and abstract factors in teaching to the exclusion of other prime requisites. This could account for the significant relationship found between the teacher's self-estimation of acceptability by students and the Professional Recording ratings which were based entirely on oral and classroom sounds from audio-recordings.

C. The Minnesota Multiphasic Personality Inventory

Summary

Two self-rated scores from each teacher were obtained from this inventory; one score indicates the amount of tendency toward social introversion and the other toward social competency. The higher the score in the former case the more the teacher tends toward undesired social tendencies; in the latter case the reverse is true. Thus correlations with negative signs constitute positive relationships where comparisons are made with social competency.

Moderately high coefficients of correlation (-.567, -.255, -.464, and -5.19) were found in almost all cases between Social Competency and the composite measures of teaching success. Social Introversion scores when compared with the teacher efficiency criteria were -.471, +.776, -.031, and -.048.

Conclusion

Social introversion, as part of the M.M.P.I. does appear to be moderately significant (-.411) to the amount of success a teacher achieves. In other words, the teacher himself may indicate he is socially introverted and can still be judged adequate as a teacher by others. One particular point worth noting is that the teacher who judges himself most socially introverted seems to believe himself to be the better teacher (+.776 between the Teacher Self-Evaluation and SI score).

When the composite of many social factors measured in the M.M.P.I. are combined into the Social Competence score there seems to be good internal consistency. A moderate degree of relationship appears to exist between teacher success and how well the teachers judge he is socially competent.

5. Are the teaching staff and student group's concepts of the teacher's social acceptability related to efficiency in teaching?

A. Teacher-Teacher Social Distance Scale (Staff Rating)

Summary

Moderately high relationships (+.431, +.581 and +.352) were found to exist between the staff's rating of the individual teacher for social acceptability and measures of teaching success except with the teacher's personal estimate (-.088) of his own teaching.

Conclusion

The degree to which a teaching staff in a school socially accepts each one of its individual members has a definite relationship to the performance of each person on the staff. Inasmuch as there is slight negative relationship between the teacher's rating of his own teaching and the staff's social acceptance of him, it is concluded that the individual teacher and the teaching staff do not consider the same aspects, especially in area of social acceptance, to be of equal value in attaining success in teaching.

B. Student-Teacher Social Distance Scale (Student Rating)

Summary

Significant coefficients of correlation (+.702, +.750, and +.369) were found to exist, in all but one case (+.069, Teacher Self-Evaluation and S-TSDS), between the Student-Teacher Social Distance Scale ratings and teaching success measures.

Conclusion

The degree to which the students in the teacher's classes accept him socially is to a considerable extent related to his effectiveness as a teacher. It is further evident that the teacher, in appraising his own professional acts, does not include social acceptance by his students as one of the requisites to teaching success, or else the instruments used are not refined enough to measure the relationship as it exists.

6. If there exists a high or low degree of social acceptability by the group/s for the individual teacher, does there also exist a high or low degree of personally judged social adjustment and success in teaching?

Summary

A coefficient of correlation of -.255 (positive relationship) was obtained between the self-rated teaching evaluation and self-rated social competency while a +.321 relationship was found existing between the self-rated teaching evaluation and the subject's self-rating on the Social Distance Scale. Coefficients of correlation of +.069 and -.086 were obtained between the particular class and staff ratings of the teacher on the Social Distance Scale and his self-evaluation of teaching.

Conclusion

The degree to which a teacher is successful in his verbal communication in a classroom apparently has definite bearing on his effectiveness as a teacher. The social competencies of oral interaction such as eliciting willing responses from students by constructive and encouraging comments; the degree to which cooperation is offered in class planning, activities, evaluation, and total learning process; the ability to take a joke and joining in with the class in a good laugh; being self-controlled in the midst of conflicting demands and never sounding angry, etc., are pertinent to the success of the teacher.

The relatively low relationship (+.171) existing between the Teacher Self-Evaluation Rating Scale and the Recording Ratings (Social Competence) apparently points out that either the teacher does not consider, or does not have the opportunity to consider because of limitations of instruments, oral social competency as important to teacher efficiency as the other teaching criteria appear to possess.

8. Do there exist mean differences between the most successful teachers and the least successful teachers, as evaluated by the various criteria of teaching success, in the measures of social competence?

Summary

1) The upper fourteen and lower fourteen groups as found by the Composite Score of Supervisory Ratings had mean differences in which the upper group appears higher in all measures of social competence.

2) It can be observed that the upper group of most successful teachers is consistently more socially competent than the group of least successful teachers where the criterion of Pupil Reactions to Instruction was used.

3) Where the criterion of Teacher Self-Evaluation was used to select the upper and lower groups, the fourteen most successful teachers have higher mean differences in five Social Competence measures. The lower group of teachers rate higher in Teacher-Teacher S. D. S. (Staff) measure, and no mean exists between the two groups in the Teacher-Student S. D. S. (Subject).

4) Higher ratings in five measures of Social Competence are to be noted for the upper group of fourteen teachers where the criterion of selection of Upper and Lower groups was based on the Recording Ratings (Professional Competencies). In two of the measures of Social Competence, the Teacher-Teacher S. D. S. (Subject) and the M. M. P. I. (SI score), the group of fourteen least successful teachers rate higher.

5) Over-all comparisons of the fourteen most successful teachers and the fourteen least successful teachers on all seven measures of Social Competence are made, and for all the criteria of teaching success. The comparisons are made between pairs of teachers, comparing the first most successful and the least successful teacher followed by the next most successful teacher and the next least successful teacher, etc. Data shows that the upper groups of teachers were higher in all Social Competence measures (227 comparisons) by twice the amount as the lower groups of teachers. Both upper and lower groups were equal in 55 comparisons.

Conclusion

There exist definite mean differences between the fourteen most successful and fourteen least successful teachers in twenty-four of the twenty-eight mean difference scores in measures of Social Competence.

The upper group appears to be higher in nearly all cases with four exceptions. These exceptions can probably be accounted for by the fact that in each case the measures involved, either the particular teaching criteria or the social competence criteria, called for self-evaluation by teacher and this study tends to point out that self-evaluations are the least consistent and reliable of the measures used.

It should be noted that the fourteen each composing the upper and lower groups were different persons or occupied different ranks in the respective groups for each of the various criteria of success in teaching.

9. Is the individual teacher who is judged as either most or least successful in teaching also rated correspondingly high or low in social competence as found in the modified sociograms?

Summary

Teacher numbers 11, 14, and 20 were judged to be the three most successful teachers in the investigation, while teacher numbers 2, 45, and 47 appear to be the least successful. Modified sociograms showing these teachers' social competence in relation to their teaching colleagues and one class of their students are found in section III.

Tables VII and VIII, Section III, contain data also showing the three most and least successful teachers compared in percent for first, second, and third social competence choices received from the staff and one class of their students. The average grand total of staff social competence choices received by the three most successful teachers was 10.4% whereas the three least successful teachers received only 0.98%. The average grand total of student social competence choices received by the three most successful teachers was 12.9% and the three least successful teachers received 7.4%.

Conclusion

The most successful teachers appear to be more socially competent than the least successful teachers according to both their colleagues and students. However, the most successful teacher is not necessarily the most socially competent person in his teaching staff according to the modified sociograms made from ratings of colleagues and students. To what extent the more successful teacher can be less socially competent than others in his group is not known and to what extent the least successful teacher is socially competent is not ascertained in this study.

10. Is the variable of length of service in teaching a significant factor in the relationship between social acceptance and success in teaching?

Summary

In-service teaching has been the major occupation and source of income from their graduation date for all of the people with exception of teacher numbers 7, 9, 29, 30, 42, 45, and 36. Each of them has been away from teaching for a period of time up to a maximum of four years between their date of college graduation and 1952. The reasons given for this were military service, illness, other occupations, etc. The maximum length of in-service teaching any of the teachers has done is twelve years and all of the people vary from this maximum down to being first-year teachers. (Table III, section II, shows this more clearly.) There is no general pattern to be found in comparing length of in-service teaching with success in teaching. Both a successful and unsuccessful teacher can equally have one, two, three, five, ten, etc., years of experience and one can be judged effective while the other is judged ineffective.

Length of in-service teaching was found to be a negligible factor when compared with the social acceptability of the teacher and no apparent relationship can be established between these two factors.

Conclusion

Length of teaching service does not appear to be a significant factor in the relationship between teacher efficiency and social acceptance. Teachers of varying lengths of service fall both in the most successful and least successful groups without any seeming pattern or apparent reason. Teachers who have an equal number of years in service are found to be designated with equal alacrity as both most successful and least successful regardless of the same time lengths of experience.

The three most successful teachers, numbers 11, 14, and 20, who are socially well-accepted by their colleagues graduated from the University of Wisconsin in 1948, 1948, and 1951, respectively. The least successful teachers, numbers 2, 45, and 47, who are not socially well-accepted graduated in 1949, 1949, and 1952, respectively. This tends to amplify the fact that both relatively good and poor teachers may have approximately the same length of time in service and still vary widely in the degrees of teaching efficiency achieved.

General Conclusions

The investigator in this study has attempted a relatively new sort of evaluation of the teacher from which an educational researcher and practitioner may be enabled to obtain further measures of his professional and social competencies. These competencies especially emphasize the social aspects relevant to how the teacher conceives of himself and his role in the school, and how colleagues, students, and other persons appraise him in that role as teacher.

The basic question which was raised earlier in the study was what constitutes an adequate criteria of teacher efficiency. It was wondered whether self-evaluations, student ratings, expert ratings, and the relationships found between them provide adequate measures of the teacher's effectiveness, or whether these techniques (used in conjunction with modified sociometric techniques), audio-recording ratings of classroom teaching and self-rated social introversion and competency measures taken from a personality inventory might not yield further and important measures of the teacher's efficiency. The basic question and its sub-question have been answered to some extent by the results of this investigation.

The use of instruments and techniques in the area of social competency with its aspects of group acceptance, group structures, group interaction, and individual social adjustment would appear to have proven fruitful in the search for more adequate criteria of teaching success. The instruments and techniques of investigation employed provide the educational researcher, teacher training institution, school administration, and the teaching staff with means whereby a more well-rounded description of the teacher and his effectiveness may be obtained.

BIBLIOGRAPHY

1. Anderson, Harold H., and others. *Studies of Teacher's Classroom Personalities, III: Effects of Teachers' Dominative and Integrative Contacts on Childrens' Classroom Behavior*, Applied Psychology Monograph No. 11 (Stanford, Calif.: Stanford University Press, December 1946), 156 pp.
2. Anderson, Harold Milton. *A Study of Certain Criteria of Teaching Effectiveness*, Unpublished doctor's dissertation, University of Wisconsin, 1952.
3. Atkinson, Gretchen. “The Sociogram as an Instrument in Social Studies Teaching and Evaluation,” *Elementary School Journal*, L, No. 2 (October 1949).
4. Barr, A. S., and Emans, Lester M. “An Analysis of Teaching Rating Scales,” *Nation's Schools*, VI (September 1930), pp. 60-64.
5. Barr, A. S. *Characteristic Differences in the teaching Performance of Good and Poor Teachers of the Social Studies* (Bloomington, Ill.: Public School Publishing Co., 1929).
6. Barr, A. S., and Singer, Arthur J. “Evaluative Studies of Teacher Education,” *Journal of Teacher Education*, IV (March 1953), pp. 65-72.
7. Barr, A. S. “Measurement and Prediction of Teaching Efficiency,” *Review of Educational Research*, VIII (June 1943), pp. 218-223.
8. Barr, A. S. “Measurement of Teaching Ability,” *Review of Educational Research*, X (June 1940), pp. 182-184.
9. Barr, A. S. “The Measurement and Prediction of Teaching Efficiency: A Summary of Investigations,” *Journal of Experimental Education*, XVI (June 1948), pp. 203-283.
10. Betts, G. L. *Evaluation Through Ratings and Other Measures of Success*, Special Survey Studies, U. S. Department of Interior, Office of Education Bulletin, 1933, No. 10. National Survey of the Education of Teachers, V (Washington, D.C.: Government Printing Office, 1935), pp. 87-116.
11. Blum, Lawrence P. “Comparative Study of Students Preparing for Five Selected Professions Including Teaching,” *Journal of Experimental Education*, VI (September 1947), pp. 31-65.
12. Bonney, Merl E. “Personality Traits of Socially Successful and Socially Unsuccessful Children,” *Journal of Educational Psychology*, XXXIV (November).
13. Book, M. F. “The High School Teacher from the Pupils' Point of View,” *Pedagogic Seminary*, XII (September 1905), pp. 239-288.
14. Boyce, A. C. “A Method for Guiding and Controlling the Judging of Teaching Efficiency,” *School Review Monographs*, No. 6 (Chicago: University of Chicago, 1915) pp. 71-82.
15. Brookover, Wilbur B. “The Relation of Social Factors to Teaching Ability,” *Journal of Experimental Education*, XXXVIII (May 1945), pp. 691-696.
16. Buros, Oscar K. *The Third Mental Measurement Yearbook* (New Brunswick, N. J.: Rutgers University Press, 1949).
17. Charters, W. W. and Waples, D. *The Commonwealth Teacher-Training Study* (Chicago: University of Chicago Press, 1929).
18. Cook, Walter W., and Leeds, Carroll H. “Measuring the Teaching Personality,” *Educational and Psychological Measurements*, VII (May 1947), pp. 399-410.
19. Criswell, Joan H. “Social Structure Revealed in a Sociometric Retest,” *Sociometry*, IV (November 1939), p. 4.
20. Domas, Simon, and Tiedeman, David V. Teacher Competence: An Annotated Bibliography,” *Journal of Experimental Education*, XIX (December 1950), pp. 101-218.
21. Ellis, A. “The Validity of Personality Questionnaires,” *Psychological Bulletin*, XLIII (1946), pp. 340-385.
22. Gotham, R. E. “Personality and Teaching Efficiency,” *Journal of Experimental Education*, XIV (December 1945), pp. 157-165.
23. Hart, J. W. *Teachers and Teaching* (New York: Macmillan Co., 1934), 285 pp.
24. Hill, George E. “Teachers' Instructional Difficulties: A Review of Research,” *Journal of Educational Research*, XXXVII (April 1944), pp. 602-615.
25. Jennings, Helen H. “Sociometry,” *Encyclopedia of Psychology* (New York: Philosophical Library, Inc., 1946), pp. 874-877.
26. Knight, F. B. *Qualities Related to Success in Teaching*, Contributions to Education, No. 120 (New York: Teachers College, Columbia University, 1922).
27. Knox, William B. *A Study of the Relationships of Certain Environmental Factors to Teaching Success*. Unpublished Ph. D.

28. Krech, David and Crutchfield, Richard S. Theory and Problems of Social Psychology (New York: McGraw-Hill Book Co., 1948).
29. Kuhlen, R., and Bretsch, H. "Sociometric Status and Personal Problems of Adolescents," Sociometry, X (1947).
30. Lins, Joseph L. "The Prediction of Teaching Efficiency," Journal of Experimental Education, XV (September 1946), pp. 2-60.
31. Lough, O. M. "Teachers College Students and Minnesota Multiphasic Personality Inventory," Journal of Applied Psychology, XXX (June 1946), p. 241.
32. Lundberg, George A. "Social Interaction Patterns in a Village," Sociometry, I (January-April 1938).
33. Mason, Harry M. "Effects of Attitudes of High School Teachers of Social Studies Upon Attitudes of Their Pupils," in Further Studies in Attitudes, Series V (Lafayette, Ind.: Purdue University, 1942).
34. Montross, Harold Wesley. "Temperament and Teaching Success," Journal of Experimental Education, XXIII (September 1954), pp. 73-97.
35. Murphy, Gardner. "The Nature of Man," Main Currents in Modern Thought, IX, No. 2 (September 1952), pp. 44-52.
36. Nance, R. D. "Masculinity-Femininity in Teachers," Journal of Educational Research, XLII (May 1949), pp. 658-666.
37. Richards, F. W. Modern Clinical Psychology (Cambridge, Mass.: Harvard University Press, 1944).
38. Rostker, Leon E. "The Measurement and Prediction of Teaching Ability," School and Society, LI (January 6, 1940), pp. 30-32.
39. Ruediger, W. C., and Strayer, C. D. "The Qualities of Merit in Teachers," Journal of Educational Psychology, I (1910), pp. 272-278.
40. Schmid, John Jr. "Factor Analyses of Prospective Teachers' Differences," Journal of Experimental Education, XVIII (June 1950), pp. 289-319.
41. Simon, Donald L. "Personal Reasons for the Dismissal of Teachers in Smaller Schools," Journal of Educational Research, XXIX (April 1936), pp. 585-588.
42. Witty, Paul A. Evaluation of Studies of the Effective Teacher, in Official Report of the American Education Research Association (Washington, D. C.: American Education Research Association of the NEA, 1948), pp. 198-204.
43. Wrightstone, J. Wayne, and others. "An Application of Sociometric Techniques to School Personnel," Journal of Experimental Education, XX (March 1952), pp. 301-304.
44. How to Construct a Sociogram, Horace Mann-Lincoln School Experimentation (New York: Bureau of Publications, Teachers College, Columbia University, 1947).

TEACHER ATTITUDES TOWARD CURRICULUM CHANGE: THE EFFECT OF PERSONAL RELATIONSHIPS ON THE DEGREE OF FAVORABLENESS

EVELYN I. BANNING*
Harvard University

Introduction

IT HAS LONG been recognized that the attitudes of teachers are central to curriculum change (6:629), and that to change the curriculum one must change the people that operate it. To understand teachers as professional workers, Miel (41) indicated the importance of knowing the teacher culture, with its characteristic strengths and weaknesses, for determining the mode of change. On the principles that the attitudes of teachers have their anchorage in the group to which they belong and that curriculum change is a social process developing out of the situation in which teachers are working, the writer formulated the underlying hypothesis of this research; namely, that interaction within the teaching situation determines the favorableness and/or unfavorableness of teacher attitudes.

The present experimental study of social attitudes originated in an attempt to determine teacher attitudes toward curriculum change at the local community level. The research is limited to the attitudes of sixty-five junior high school classroom teachers in a New England community of approximately 50,000 population. The testing locale was selected because the local administrative and teaching staff had been actively involved for several months in junior high school curriculum revision.

Three broad implications for curriculum development were considered preliminary to the actual research:

1. The administrative and supervisory staff as status leaders might have a means and be in a better position to facilitate curriculum development.
2. Clearer insight into the teachers' own feelings about curriculum change might lead to increased teacher competency.
3. The investigation of teacher attitude within the particular organization was likely to determine readiness of change.

Purpose of the Study

The purpose of this exploratory study was two-fold: (1) to determine, insofar as possible, the favorableness and/or unfavorableness of the attitudes of the sixty-five junior high school teachers; and (2) to relate the degree of favorableness to selected variables within the teaching situation to discover to what extent these variables were positively related to favorableness and to what extent they were the causes of resistance to change.

Definitions

For reasons of objective research, the term attitude is defined in this study as a "delimited totality of behavior" (60:51), not as something that predicts behavior. It is used interchangeably with opinion, with the distinction that attitudes include opinions (verbal responses) and consist of both verbal and non-verbal responses. Since the writer is dealing with verbal "behavior," frequently classified as opinion, this study is restricted to verbal responses observable through questionnaire and interviewing. The "behavior" of interest to the investigation is the response of individuals to questions which belong in the universe of curriculum change, the concept being investigated. This concept does not involve a specific change nor proposed curriculum organization toward which change is directed. Curriculum is thought of as a process, not as a body of content; curriculum change, viewed as a type of social change becomes, therfore, largely a matter of discovering and applying better procedures for improved learning experiences for pupils.

Methodology

An attitude scale was first drafted by the writer based on the literature in the field of curriculum change, together with as complete a knowledge and understanding of the local teaching situation as possible through direct observations, meetings, informal talks, and study of the organization of each junior high school unit and its function within the system. The scale contained questions about which the teachers, already involved in curriculum revision, might be expected to have attitudes and opinions. The thirty-one items covered the following points: estimation of the present curriculum, administrative leadership, teacher responsibility, methods, source and result of change, interest in curriculum planning, freedom for classroom teaching, and effect of the curriculum study on the total school system. To each statement on the attitude scale was added the question of intensity of feeling: "How strongly do you feel about this?" with answer categories of "Not at all strongly", "Not so strongly", "Fairly strongly", "Very strongly", and "No answer". Repeating this question after each content item yielded a series of intensity answers.

---

*Graduate School of Education

For measuring the interaction of teachers within the social structure of the six junior high schools in the school system, factors within the "teacher culture" were selected as independent variables. Factors in the situation to be related to the degree of favorableness or unfavorableness of teacher attitudes in curriculum change were the variables of the social structure of each of the six junior high school units. These variables included group relationships among members of the school, teachers, pupils, and administrators; the status of teachers in the school and in the community; teacher expectations with respect to curriculum values, extent of acceptance of teachers by others as participating members of the total group; and opportunity for suggesting curriculum changes. In addition to these independent variables for measuring the situation, the variables of age, sex, and professional training and experience were selected for determining their influence, if any, on the relationships existing within the teaching groups.

The administration of the Teacher Attitude Scale to the sixty-five junior high school teachers was followed within four months by an interview with each teacher in order to gain deeper insight into attitudes and to uncover some of the causes and reasons that lay behind expressed attitudes.

Assumptions Underlying the Study

Several assumptions underlie this present study of teacher attitudes as measured both by an attitude scale and by personal interviewing. It is assumed that what teachers say they believe is actually what they believe. The attitude scale was used in a situation in which teachers could be expected to tell the truth about their own convictions and opinions. It is further assumed that the individual responses in the interviews were free and complete expressions of attitudes. The validity of verbal responses is the basic assumption of the measuring devices used in the study.

Murphy and his associates support the validity of test content as follows:

> Actions are no more inherently 'valid' in the first place, than words. The following remarks seem to us patently true: Actions are frequently designed to distort or conceal 'true' attitude quite as fully as verbal behavior.... All behavior is subject to modification in the process of execution from considerations of courtesy, expediency, or other social pressures. And it is furthermore apparent that when verbal behavior is used to distort or conceal the 'true' attitudes, the distortion commonly conforms to everyday behavior. The reasons for concealing 'true' attitudes are the same for both verbal and 'overt' behavior. If conditions of secrecy, and preferably of anonymity, are observed, there is more reason to expect free and complete expression of attitudes through words, thus freed from social pressures, than from behaviors which are open to all beholders.... And, finally, it may be observed that a man's categorical agreement or disagreement with a rather strongly stated opinion about the Chinese, or Jews, or Rotarians, or Communists, is in everyday life regarded (if the man is sincere) as a significant part of his behavior. There seems to be no reason why this behavior should suddenly become non-significant when it is made the subject of careful inquiry, particularly if motives of insincerity are reduced to a minimum. (44:912)

Whether or not a person acts in accordance with his attitudes is a question apart from its measurement and definition. That a teacher favors curriculum change and endorses favorable opinions concerning it does not imply that he will work to bring about change. No one-to-one correspondence between verbally expressed attitudes and actual behavior is indicated or implied. Since the study is limited to statements of what the individual claims his attitude to be, it is recognized that the methods are most likely to be valid if teachers have no motive to conceal attitudes.

Ranking of Respondents on Scale

Both the content questions and the intensity questions were scored according to assigned weights. The categories of the attitude (content) questions were given weights beginning with zero. In each question, the higher weights were assigned to categories judged to express a more favorable attitude toward change.

To test the scalability of the universe of curriculum change, the Cornell technique for scalogram analysis was used (24). The technique, a procedure for testing the hypothesis that the universe is a scale for a given population, was first developed for teaching purposes at Cornell.

> The universe is said to be scalable for the population if it is possible to rank the people from high to low in such a fashion that from a person's rank alone we can reproduce his response to each of the items in a simple fashion.'' (24:249)

The responses of the total sixty-five classroom teachers in the junior high school units were observed on the thirty-one items covering the general content of attitudes toward curriculum change. Two approximations, or trials, were made to reject or accept the hypothesis of scalability. The first trial indicated an error of reproducibility, considerably more than 15 percent of the 2015 responses (2015 = 31 × 65), the number of questions times the number of people.

To reduce error, categories were combined by first examining the overlapping of the scores within the columns of each question. When categories intertwined, they were combined to minimize the error of reproducibility for the combinations. On the basis of the combined categories, a second trial rank order was established by reassigning weights. Each scale was rescored and new tables made.

The error of reproducibility was much smaller than in the first trial. In question 1, six of the top seven scores were made by six of the top seventeen people as opposed to two in the first trial. However, the remaining scores were scattered the length of the column. In each column, cutting points were placed so as to minimize the errors. The total number of errors in the whole table was 505, which was 25 percent of all the responses. In terms of combined categories, this universe was found to have only 75 percent accuracy for reproducing a response from a person's rank order. Since apparently there was not a single ranking possible in the total content covered by the scale, it seemed worthwhile to break the content down into sub-areas and analyze these sub-areas separately by scalogram analysis. Thus the relationships of a sub-area to outside variables could be studied if any of these areas were found scalable.

The totality, or universe, was therefore divided into five sub-areas, made up of five or six questions each, and scaled according to the Cornell technique used for the total scale. The five sub-areas selected for scaling were as follows:

1. Attitude toward teacher responsibility in curriculum change
2. Satisfaction with the present curriculum
3. Attitude toward effecting curriculum change
4. Attitude toward the administrative leadership in the curriculum study
5. Attitude toward the value of curriculum planning

Selection of the Scale

Since the error of reproducibility of the scaled sub-areas on attitude toward teacher responsibility and attitude on methods of effecting curriculum change was greater than .10 (.19 and .13 respectively), they were rejected as scales for measuring the favorableness of attitude toward change. Although the three remaining sub-areas scaled with a coefficient of reproducibility higher than .90, the six items of the sub-area on the value of curriculum planning were judged by the writer as more indicative of attitude toward change in the teaching situation in the junior high schools than the items on satisfaction with the present curriculum and on administrative leadership. This selection was made, as was that of the original set of thirty-one items, by the subjective judgment of the writer with full consideration of the given population and the time sequence.

For several months prior to the administration of the original scale, the teachers had been actively engaged in curriculum revision through group meetings. In addition, the administrative staff had directed the thinking of the teachers, through an unsigned questionnaire, concerning the principles and practices of the curriculum, broadly conceived. Through these experiences, the junior high school teachers had been considering curriculum change as an on-going process of change with no preconceived specifics and no proposed curriculum organization toward which change was to be directed. Consequently, it was the writer's contention that the general content items on the opinions of teachers about the value of curriculum change as a process would measure the favorableness of teacher attitude within the local situation more effectively than the other items in the two scalable sub-areas.

Since the teachers were already involved in curriculum revision, it was unlikely that teachers would indicate a completely negative response on the question: "All things considered, do you think the present junior high school curriculum is about as satisfactory as possible?" Actually the response "Completely Satisfactory" was unused, thus necessitating re-weighting of the categories by its omission. It was also apparent to the writer that the content questions on satisfaction toward the present curriculum limited the responses to specifics such as choice of subjects, courses, and facilities and equipment. Therefore, although the sub-area scaled with a .91 accuracy, it appeared too restrictive in subject coverage to measure the concept of curriculum change as a process.

The sub-area on attitude toward administrative leadership was likewise considered too limited in content to measure attitude toward change, although the writer recognized that teacher attitude toward leadership and administrative policies might be related to change and might be a factor in determining change.

Of the original thirty-one items, the six items that covered the fundamental concept of curriculum change as defined previously as a process, not as a method, course of study, or body of content were scaled and found usable. Although these items were selected initially by reason of subject content, the sub-area scaled somewhat higher than either of the other two scalable sub-areas. It not only had the highest coefficient of reproducibility (.93), but the range of marginal frequencies was evenly distributed over all the categories, and the scatter of errors in each item was random. Thus, it conformed best to the scaling features of the Cornell technique.

Since it was believed that the two scalable sub-areas on satisfaction with the present curriculum and attitude toward administrative leadership might actually be independent of the selected scale, they were related to the sub-area on the value of curriculum planning. The relationship between the attitude on satisfaction with the value of curriculum planning was as follows:

Attitude Toward Value

	U	F	
Dissatisfied	11	33	44
Satisfied	12	9	21
	23	42	65

The hypothesis that the value of curriculum planning is independent of satisfaction was tested by means of chi-square. Since the probability of occurrence of a chi-square of 5.09 is less than .05 if the hypothesis is true, the hypothesis was rejected. Therefore, the data indicate that attitude toward the value of curriculum planning and satisfaction with the present curriculum are related to some degree. Sixty-eight percent of all sixty-five responses showed agreement between the attitudes expressed on the two scales.

The relationship between the attitude toward administrative leadership and the value of curriculum planning was as follows:

Attitude Toward Value

	U	F	
Attitude Toward Leadership	7	29	36
	16	13	29
	23	42	65

The hypothesis that attitude toward administrative leadership is independent of the value of curriculum planning was also tested by means of chi-square. Since the probability of occurrence of chi-square of 7.47 is less than .01 if the hypothesis is true, the hypothesis was rejected. Therefore, the data indicate that the attitudes expressed on the two scales related to some degree. Sixty-eight percent of all sixty-five responses showed agreement.

Thus there is evidence that in general the teachers who tend to express a favorable attitude toward change also tend to express a favorable attitude on administrative leadership and dissatisfaction with the present curriculum, while those who express an unfavorable attitude toward change tend to express an equally unfavorable attitude on administrative leadership and satisfaction with the present curriculum. For this reason, only the scale on the value of curriculum planning was related to the interview variables, since this independent variable was the major concern of the study and since similar findings might be expected of the other scales in relation to the interview variables.

Table I indicates the questions and frequencies of responses for the six items used in the scale. The scale analysis for this scalogram is shown in Table II.

Intensity Function

To determine a cutting point that would be meaningful with respect to the whole scale continuum, the two-part technique of the intensity function was used. The fold-over technique had been previously rejected, since Stouffer (60:253) stated that it would seem safer to use separate intensity questions, which would call for distinct and separate decisions from the respondent. In the two part technique, intensity scores are experimentally independent of the content scores.

Use of the intensity function obviated the problem of question bias, in that it would yield the same proportion of the group as favorable

TABLE I

ATTITUDE TOWARD THE VALUE OF CURRICULUM PLANNING

Scale Item	Number of Question	Frequencies of Response
1	7	Do you usually feel that what you are doing in the curriculum study is worthwhile or not? Favorable 45 Unfavorable 20
2	9	How interested are you in your part in curriculum planning? Favorable 43 Unfavorable 22
3	20	Do many of the pre-planning activities in the curriculum study seem important? Favorable 9 Unfavorable 56
4	21	How do you feel about the importance of curriculum study as compared with other work you might be doing? Favorable 35 Unfavorable 30
5	27	Do you have time to be active in the curriculum study and still prepare for your teaching to your own satisfaction? Favorable 17 Unfavorable 48
6	31	Do you feel that the present curriculum study for the junior high schools has value for the total program of the entire school system? Favorable 60 Unfavorable 5

TABLE II

SCALE ANALYSIS FOR TEACHER ATTITUDE SCALE
SECOND TRIAL: CONTENT

	1		2		3		4		5		6	
Score	2	0	2	0	2	0	2	0	2	0	2	0
12	x		x		x		x		x		x	
12	x		x		x		x		x		x	
12	x		x		x		x		x		x	
12	x		x		x		x		x		x	
10	x			x	x		x		x		x	
10	x		x			x	x		x		x	
10	x		x			x	x		x		x	
10	x		x			x	x		x		x	
10	x		x			x	x		x		x	
10	x		x			x	x		x		x	
10	x		x			x	x		x		x	
10	x		x			x	x		x		x	
10	x		x			x	x		x		x	
8	x		x			x		x	x		x	
8	x		x			x		x	x		x	
8	x		x			x	x			x	x	
8	x		x			x	x			x	x	
8	x		x			x	x			x	x	
8	x		x			x	x			x	x	
8	x		x			x	x			x	x	
8	x		x			x	x			x	x	
8	x		x			x	x			x	x	
8	x		x		x			x		x	x	
8	x		x			x	x			x	x	
8	x		x			x	x			x	x	
8	x		x		x			x		x	x	
8	x		x			x	x			x	x	
8	x		x			x	x			x	x	
8	x		x		x		x			x		x
6		x	x			x	x			x	x	
6		x	x			x	x			x	x	
6		x		x		x	x		x		x	
6	x			x		x	x			x	x	
6		x	x			x	x			x	x	
6	x		x			x		x		x	x	
6	x		x			x		x		x	x	
6	x		x			x		x		x	x	
6	x		x			x		x		x	x	
6	x		x			x		x		x	x	
6	x		x			x		x		x	x	
6	x		x			x		x		x	x	
6		x	x			x		x	x		x	
4		x	x			x		x		x	x	
4	x			x	x			x		x	x	
4	x			x		x		x		x	x	
4	x			x		x		x		x	x	
4	x			x		x		x		x	x	
4	x			x		x		x		x	x	
4	x			x		x		x		x	x	
4	x			x		x		x		x	x	
4	x			x		x		x		x	x	

TABLE II (Continued)

	1		2		3		4		5		6	
Score	2	0	2	0	2	0	2	0	2	0	2	0
4		x		x		x	x			x	x	
4		x	x			x		x		x	x	
4		x		x		x	x			x	x	
4		x	x			x		x		x	x	
4		x		x		x	x			x	x	
4		x		x		x	x			x	x	
2		x		x		x		x		x	x	
2		x		x		x		x		x	x	
2		x		x		x		x		x	x	
2		x		x		x		x		x	x	
2		x				x		x		x		x
2		x	x	x		x	x			x		x
0		x		x		x		x		x		x
0		x		x		x		x		x		x
Frequency	45	20	43	22	9	56	35	30	17	48	60	5

TABLE III

PUPIL-TEACHER RELATIONSHIP*

Attitude Toward Change	Handling Individual Disturbances			Number of Responses
	Through Individual Help	Through Pupil Assistance	By Class Recognition-Teacher Authority	
Unfavorable	24	22	52	23
Favorable	76	78	48	42
	100	100	100	
Total Responses	29	9	27	65

*Data reported in percentages.

TABLE IV

ADMINISTRATOR-TEACHER RELATIONSHIP*

Attitude Toward Change	Channels of Communication				
	Outside Consultants	Leaders Within System	Individual Teachers	Superintendent to Principal	Number of Responses
Unfavorable	0	29	33	44	23
Favorable	100	71	67	56	42
	100	100	100	100	
Total Responses	2	17	21	28	65

*Data reported in percentages.

TABLE V

COMMUNITY ACTIVITY*

Attitude Toward Change	Very Active	Fairly Active	Not Very Active	Not Active	Number of Responses
Unfavorable	0	29	46	60	23
Favorable	100	71	54	40	42
	100	100	100	100	
Total Responses	8	24	28	5	65

*Data reported in percentages.

and unfavorable, no matter how the questions were worded, what the position of the question was, and what order was used in the presentation of the answer categories. Thus this objective method offered an invariant zero point for attitudes. It is based upon the concept that it is possible to measure the intensity of feeling which people at different scale positions hold their attitudes, that such intensity marks off a point of "neutrality." This provides a division of the group into two sections, one positive and the other negative.

To each of the thirty-one content questions on the teacher attitude scale, a second question on the strength of feeling was added, "How strongly do you feel about this?" The five categories including degrees of strength of feeling ("Very strongly," "Fairly strongly," "Not so strongly," and "Not at all strongly," and "No answer") were assigned weights of 4, 3, 2, 1, 0, in that order. The first trial of the intensity scores indicated that dichotomization of these categories might give a quasi-scale of intensity. New weights were assigned, 2 for "Very strongly" and 0 for all other responses (60:253). The intensity scores ranged from 0 to 12 depending upon the number of answers about which each teacher felt very strongly.

The lowest point, a very definite zero point, fell at 24 percent. The sharpness of the curve indicated that there were few teachers who were "neutral" in their attitudes toward curriculum planning for change in the junior high schools. The ratio of approval to disapproval, or favorableness to unfavorableness, was approximately one to three. Such a very sharp curve showed a clear distinction between being positive and being negative. The low point was in the negative region of the content continuum, showing more teachers in favor of curriculum change than opposed to its planning.

The curve obtained for an intensity function on the sixty-five cases used seems to differentiate reliably between positives and negatives, to the degree that it is safe to say that, according to the sample of junior high school teachers on the teacher attitude scale, 65 percent held a favorable attitude toward the value of planning for curriculum change, and 35 percent held unfavorable attitude toward such curriculum planning.

Following twelve identification codes, the interviews yielded forty-seven items dealing with factors in the teaching situation. Each variable was correlated with the attitude score of the Teacher Attitude Scale Toward the Value of Curriculum Planning (see Table I). Owing to the limited number of categories for the small population (sixty-five teachers), four-fold tables were used in correlation. The chi-square test of independence was first applied to each 2/2 table to determine, in advance of measuring the degree of correlation, the significance of the correlation as a chance departure from zero or no relationship. To the tables that showed a probability of .05 or smaller, the Yates correction formula for continuity was applied. The degree of correlation was then measured by tetrachoric correlations, computed by the manual[1] for the corrected chi-squares with a probability not larger than .05.

Group Relations

Within the teaching organization itself, three relationships were tested: pupil-teacher, administrator-teacher, and teacher-teacher relationships. Responses to interviewing were the determining factor for these group relations. Pupil-teacher relationship was judged by the introductory interview questions, since it had been noted in the pre-test of the interview that teachers generally responded with less restraint and more rapport when the initial question concerned pupil relations with other pupils.

When chi-squares were applied to the interview items concerning pupil-teacher relationships, it was found that one item differentiated between teachers favorable and unfavorable toward change. Item 3: "In general, how do you handle pupils when they want to do something you feel is disturbing to the group?" revealed a correlation between attitude toward curriculum change and teacher-pupil relationships with classroom activities.

Teachers who indicated that they handled each adjustment problem individually through personal guidance or help in class time were generally more favorable in their attitude scores. This situation was also true of the teachers who recognized the status of the pupil among his peers and arranged a classroom working relationship advantageous to both the individual and the class group.

With a chi-square of 4.31 the hypothesis that no relationship exists between pupil-teacher relationships in classroom activities and attitude toward curriculum change is untenable.

The significant pupil-teacher relationships concerned the attitudes of teachers toward the handling of pupil-adjustment problems in the classroom. Favorableness toward change was related in a positive degree to the tone teachers maintained in dealing with individual pupils who disturbed the planned group activity. Consideration of the disturbance as an individual matter to be adjusted according to the benefit and educational growth of the pupil rated teachers more highly favorable than other approaches. In particular, the teachers who felt that class disturbances could be automatically handled by dismissal from the classroom, by removal of privileges, or by reprimand before the class in order to retain teacher authority were generally more unfavorable toward curriculum change.

---

[1] Lenore Chesire, Milton Saffir, and L. L. Thurstone, Computing Diagrams for the Tetrachoric Correlation Coefficient (Chicago: University of Chicago, 1933).

Teacher relationship to administration showed a significant difference in respect to the determining of curriculum policies for change. To the question: "Do you feel that teachers have a real part in determining the policies of the curriculum?" teachers indicated the channels of communication by which they felt the policies were determined. These responses differentiated to a real degree the favorableness and unfavorableness of teacher attitudes toward change. The teachers who felt that communication was a formal channel from superintendent and principals to the teaching staff with little return communication evidenced no difference, but teachers who recognized individual responsibility and assistance of leaders, either outside consultants, or directors and supervisors within the system, were more favorable toward change.

Items 1, 2, 3 were combined into one variable of non-formal channels of communication in order to relate formal and informal channels to favorableness of attitude toward change. The relationship was studied by means of the chi-square test. The obtained chi-square of 6.16 showed that there was a probability of .01 that a relationship existed.

Favorableness of attitude toward change correlated significantly with the feeling that the teacher was a real participant in the formulation of curriculum policy, and that this individual contribution comprised not only advice in policy-making but also a recognized share in implementing policy decisions. This relationship was supported by a like positive feeling of agreement with, or approval of, the administrative leadership in the curriculum study. It seemed evident that the interaction of teachers and administrators in a working relationship determined to a considerable degree the favorableness of teacher attitude toward change.

Constellations of friends within the community and membership in civic groups showed little distinction between teachers who were favorable to change and those unfavorable. Teacher activity within the life of community, however, did indicate a somewhat significant difference.

To the extent that teachers were increasingly active, the degree of their favorableness toward change increased. Teacher responses were easily ascertainable on the item of activity, for teachers felt very strongly about their own community participation. Non-active teachers explained their reasons for inactivity or stated their feelings about the questioned value of additional out-of-school activities. In coding items for teachers who indicated the extent of participation in organized community life, the writer considered membership in four groups as average.[2]

With a chance probability of slightly less than .05, it is possible to state that some relationship exists between degree of community activity and attitude toward curriculum change.

The opinion that teachers expressed concerning their activity in the life of the community, when correlated with scores on the attitude scale, showed a significant relationship. The sense of belonging to the community as a participating member rather than membership in a particular kind or number of civic, religious, fraternal, or professional organizations showed a relationship to favorableness of change. Whether or not the teacher had contact with pupils outside the school seemed not to affect this relationship, but if teachers felt that they were active in the adult life of the community, they tended to express a more favorable attitude on the scale. All of the junior high school teachers included in this study stated they lived in the city, yet 51 percent felt that, although they were residents of the community, they did not actively participate in its civic life.

Patterns of Satisfaction

From the total responses of the interviews, it was possible to discover underlying patterns of dissatisfaction or satisfaction of teachers toward the present junior high school curriculum. In correlating the coded variables with these patterns of dissatisfaction or satisfaction with existing conditions, the writer found two factors that were significantly related: sex and pre-appointment training. Of the twenty men teachers interviewed, 80 percent stated a feeling of dissatisfaction with the existing curriculum. Women teachers, on the other hand, much more frequently (42 percent) expressed a feeling of satisfaction. A chi-square of 6.50 indicated that there was a probability of less than one chance in a hundred that this difference was random.

Graduates of liberal arts colleges and universities generally felt more dissatisfied with conditions than teachers who had received their undergraduate preparation at state teachers colleges. There was a marked tendency for graduates of other than state teachers colleges to feel dissatisfied toward the present curriculum and to indicate a greater readiness for change than was indicated by graduates of the teachers colleges. Degree status, or course credits and degrees received since appointment, did not affect the ratings nor show a significant relationship with feelings of dissatisfaction, indifference, or satisfaction.

---

[2] Florence Greenhoe, "The Community Contacts and Participation of 9,122 Public-School Teachers Selected as a National Sample," School and Society, (October 14, 1939), pp. 510-512.

The interview responses also yielded evidence that willingness to be active in curriculum change was expressed by as many teachers who were satisfied with conditions as teachers who were dissatisfied. Teachers who indicated such willingness were generally favorable toward the values of integration. Highly significant was the correlation of willingness to exert effort for curriculum change with the favorableness of attitude toward integration for classroom experiences. Teachers who recognized the importance of unifying learning experiences for pupils were differentiated from those who had conveyed the opinion that subject content or a particular subject, should be taught as a separate area. Teachers ready to take an active part in effecting change in the curriculum generally felt that much more value accrued to the pupil through experiences and activities that cut across subject lines than through separate subject matter. Among the teachers interviewed, readiness for curriculum change appeared to be related to the conviction that educative experiences should be included in the curriculum on the basis of their integrative values for junior high school pupils, and that no subject should be taught in isolation, apart from the interrelating whole.

Results of the Study

The results of the findings indicate the importance of the ability of teachers to communicate and the necessity for clear channels of communication, both within the teaching units and from teachers to administrators and from administrators to teachers. More essential than that the direct line of communication from teacher-to-principal-to-superintendent be used is that a two-way process shall be continuously maintained. Teachers need to understand the decisions made by the administrators for promoting changes in the school curriculum or school organization if they are to accept responsibility for making their individual contributions. The way teachers feel about a decision is more important than the decision itself. Without the favorable attitude of teachers and their active personal interest as cooperative members of the group, change in curriculum seems likely to be superficial or at least temporary.

Parallel to the need for formal organization with established channels of communication is that of the opportunity for teachers to share in cooperative group action. Such action necessitates a situation so clearly structured that each teacher would feel secure and free to participate in the organization. Promoting this kind of cooperative activity requires, too, the kind of atmosphere that would permit and encourage all teachers concerned in a problem to take part in solving it.

From the expression of pupil-teacher relationships in this study, the writer has been led to conclude that harmonious relationship with pupils is largely determined by the ability of teachers to think in terms of the culture—the way of life—of his pupils. Even when the relationship involved considerable conflict or occasional strong antagonism, many teachers indicated with apparent honesty that they saw no reason for treating pupils as individuals, that all children ashould have the same treatment. Many who recognized the theoretical value of experiences suited to individual children recalled their own experiences in the classroom, remembering no occasion when they had received individual attention. Expectations of the classroom were the same for all. Thus having their roots in one culture some teachers were finding it uncomfortable to adjust to another.

On the other hand, there were teachers who desired to apply the principle of individual differences but lacked the necessary support and encouragement from the supervisory or administrative staff and were therefore fearful of misunderstanding by other teachers. Unless both administration and teaching staff recognize to a greater degree the value of a positive approach in helping pupils develop cooperative skills for group living, revising the curriculum may well continue to mean mere courses of study and materials to be covered. Of more importance for curriculum change than teaching load, size of class, or free time during the school day is the classroom atmosphere in which each pupil feels he is an accepted and contributing member of the group, and in which the relationship between teachers and pupil is congenial and productive of satisfying experiences for both.

A further conclusion of this exploratory study is that in addition to the effect teacher-administrator and teacher-pupil relationships have on the attitudes of teachers toward the value of curriculum planning, the teachers' belief in the need for continuous curriculum change is closely related to a "sense of belonging" in the life of the community. Teachers who felt strongly that they contributed to the civic life of their community were likely to be favorable in their attitudes. Community status related significantly to favorableness toward curriculum change. Such a relationship indicates the importance of teachers participating responsibly in community life and needs of its people.

In summary, this study of the factors in the teaching situation that related in a positive way to attitudes toward curriculum change reveals that teachers are more favorable if they feel that they are making a contribution to the school organization as a whole and that they have had a meaningful share in policy decisions and their implementations; if their relationship with pupils in the classroom is a harmonious one; and if they are active participants in the life of the community. The findings indicate that these four factors affect the attitudes of teachers more than the factors of teaching-load, subject or grade taught, or any of the mechanical conditions of work.

The findings of this study further substantiate the central hypothesis that the degree of favorableness of teacher attitudes toward curriculum change relates significantly to the factors within the teaching situation, most specifically to the interaction of teacher and administrators, and of teachers and pupils in classroom activities. Whether this interaction determines the attitude or is instead determined by a factor yet unmeasured awaits further study and experimentation.

BIBLIOGRAPHY

1. Allport, Gordon W. "Attitudes" in Carl Murchison (ed.), Handbook of Social Psychology (Worcester, Mass.: Clark University Press, 1935), pp. 798-844.
2. Association for Supervision and Curriculum Development. Action for Curriculum Improvement, 1951 Yearbook (Washington, D.C.: National Education Association, 1951).
3. Bain, Read. "Theory and Measurement of Attitudes and Opinions," Psychological Bulletin, XXVII (1930), pp. 357-380.
4. Bain, Winifred E. "A Study of Attitudes of Teachers Toward Behavior Problems," Child Development, V (March 1934), pp. 19-35.
5. Barr, Arvil S. "Impressions, Trends, and Further Research," Journal of Experimental Education, XIV (December 1945), pp. 2-4-206.
6. Barr, Arvil S., and others. Supervision (New York: D. Appleton-Century Co., 1947).
7. Barron, Caroline K. "Points of Resistance to Curriculum Change," Progressive Education, XXVIII (May 1951), pp. 201-205.
8. Benne, Kenneth D., and Muntyan, Bozider. Human Relations in Curriculum Change: Selected Readings with Special Emphasis on Group Development, Circular series A, No. 51, Illinois Secondary-School Curriculum Program Bulletin, No. 7 (Springfield, Ill.: Office of State Superintendent of Public Instruction, June 1949).
9. Campbell, Angus. "Polling, Open Interviewing, and Problem of Interpretation," Journal of Social Issues, II (November 1946), pp. 67-71.
10. Cantril, Hadley (ed.). Gauging Public Opinion (Princeton, N. J.: Princeton University Press, 1944).
11. Cantril, Hadley, "The Intensity of an Attitude," Journal of Abnormal and Social Psychology, XLI (1946), pp. 129-135.
12. Caswell, Hollis L. Education in the Elementary School (New York: American Book Co., 1943).
13. Chapin, F. Stuart. Experimental Designs in Sociological Research (New York: Harper and Bros., 1947).
14. Clark, Elmer J. "Teacher Relations Toward Objectionable Pupil Behavior," Elementary School Journal, LI (April 1951), pp. 446-449.
15. Cralle, Robert E., and Burton, William H. "An Examination of Factors Stimulating or Depressing Teacher Morale," California Journal of Elementary Education, VII (August 1938), pp. 7-14.
16. Cunningham, Ruth, and Miel, Alice. "Frontiers in Research in the Elementary School Curriculum," Journal of Educational Research, XL (January 1947), pp. 365-372.
17. Day, Daniel. "Methods in Attitude Research," American Sociological Review, V (June 1940), pp. 395-410.
18. Droba, D. D. "Methods of Measuring Attitudes," Psychological Bulletin, XXIX (May 1932), pp. 309-323.
19. Dyer, Henry S. Observable Evidence of Prejudice in Ninth-Grade Children, Unpublished D. Ed., thesis, Harvard University, Graduate School of Education, 1941.
20. Edwards, Allen L. Experimental Design in Psychological Research (New York: Rinehart and Co., 1950).
21. Factor, Bernice. "Curriculum Change in Relation to Certain Psychological and Cultural Factors," Harvard Educational Review, XIX (October 1949), pp. 219-230.
22. Fisher, Ronald A. Statistical Methods for Research Workers. 9th ed., (London: Oliver and Boyd, 1944).
23. Guttman, Louis. "A Basis for Scaling Qualitative Data," American Sociological Review, IX (April 1944), pp. 139-150.

24. ———. "The Cornell Technique for Scale and Intensity Analysis," Educational and Psychological Measurement, VII (1947), pp. 247-279.

25. ———. "On Festinger's Evaluation of Scale Analysis," Psychological Bulletin, XLIV (1947), pp. 451-465.

26. Harap, Henry, (ed.). The Changing Curriculum (New York: D. Appleton-Century Co., 1937).

27. Harper, Manly H. Social Beliefs and Attitudes of American Educators (New York: Bureau of Publications, Teachers College, Columbia University, 1927).

28. Harris, Pickens E. The Curriculum and Cultural Change (New York: D. Appleton-Century Co., 1937).

29. Hartmann, George Wilfred. "Social Attitudes and Information of American Teachers," in W. H. Kilpatrick (ed.), The Teacher and Society (New York: D. Appleton-Century Co., 1937), pp. 174-230.

30. Hoppock, Robert. Job Satisfaction (New York: Harper and Bros., 1935).

31. Katz, Daniel. "The Measurement of Intensity" in Hadley Cantril (ed.), Gauging Public Opinion (Princeton, N. J.: Princeton University Press, 1944), pp. 51-65.

32. Katz, Daniel and Allport, F. H. Students' Attitudes: A Report of the Syracuse University Reaction Study (Syracuse, N. Y.: Craftsman Press, Inc., 1931).

33. Kroll, Abraham. "The Teacher's Influence Upon the Social Attitudes of Boys in the Twelfth Grade," Journal of Educational Psychology, XXV (1934), pp. 274-280.

34. Linton, Ralph. Study of Man (New York: D. Appleton-Century Co., 1936).

35. Lund, F. H. "The Psychology of Belief," Journal of Abnormal and Social Psychology, XX (1925-1926), pp. 193-195.

36. Manske, Arthur J. The Reflection of Teachers' Attitude in the Attitudes of Their Pupils (New York: Bureau of Publications, Teachers College, Columbia University, 1936).

37. McClusky, Howard Y., and Strayer, Floyd J. "Reactions of Teachers to the Teaching Situation: A Study of Job Satisfaction," School Review, XLVIII (October 1940), pp. 612-623.

38. McNemar, Quinn. "Opinion Attitude Methodology," Psychological Bulletin, XLIII (July 1946), pp. 289-374.

39. ———. Psychological Statistics (New York: John Wiley and Sons, Inc., 1949).

40. Miel, Alice. "Barriers to Improved Instruction," Teachers College Record, XLVI (April 1945), pp. 434-440.

41. ———. Changing the Curriculum: A Social Process (New York: D. Appleton-Century Co., 1946).

42. ———. Curriculum Change as a Deliberate Social Process in Local Communities, unpublished D. Ed., thesis, Teachers College, Columbia University, 1944.

43. Mitchell, Claude. "Do Scales for Measuring Attitudes have any Significance?" Journal of Educational Research, XLIII (February 1941), pp. 445-452.

44. Murphy, Gardner and others. Experimental Social Psychology (New York: Harper and Bros., 1937).

45. Nelson, Erland. "Attitudes," Journal of General Psychology, XXI (October 1939), pp. 367-436.

46. Peck, Leigh. "Teachers' Reports on the Problems of Unadjusted School Children," Journal of Educational Psychology, XXVI (February 1935), pp. 123-138.

47. Prescott, Daniel A. Emotion and the Educative Process: A Report of the Committee on the Relation of Emotion to the Educative Process (Washington, D. C.: American Council on Education, 1938).

48. Prall, Charles and Cushman, C. Leslie. Teacher Education in Service (Washington, D. C.: American Council on Education, 1944).

49. Remmers, H. H. (ed.). "Studies in Attitudes: A Contribution to Social-Psychological Research Methods," Purdue University Bulletin, XXV, Studies in Higher Education, No. 26, 1934.

50. Remmers, H. H. and Silance, Ella B. "Generalized Attitude Scales," Journal of Social Psychology, (August 1934), pp. 298-311.

51. Rogers, Dorothy. "The Interview Technique in Obtaining Teacher Opinion," Educational Administration and Supervision, XXXVI (November 1950), pp. 419-424.

52. Rugg, Harold. "Curriculum Making: Past and Present," Twenty-Sixth Yearbook of the National Society for the Study of Education, Part I (Bloomington, Ill.: Public School Publishing Co., 1926).

53. Schank, R. L. "A Study of a Community and Its Group and Institutions Conceived as Behavior of Individuals," Psychological Monographs, XLIII, No. 195 (1932).

54. Sherif, Muzafer and Cantril, Hadley. The Psychology of Ego-Involvements: Social Attitudes and Indentifications (New York: John Wiley and Sons, Inc., 1947).

55. Suchman, Edward A. "The Logic of Scale Construction," Educational and Psychological Measurement, X (Spring, 1950), pp. 79-93.

56. Sherman, Mandel. "Theories and Measurement of Attitude," Child Development,

III (March 1932), pp. 15-28.

57. Smith, B. Othanel and others. Fundamentals of Curriculum Development (New York: World Book Co., 1950).

58. Southworth, A. E. "Attitudes Favorable to Change," National Association of Secondary-School Principals Bulletin, XXXIV (December 1950), pp. 134-141.

59. Stagner, Ross "Attitudes," in W. S. Monroe (ed.), Encyclopedia of Educational Research, Rev. Ed. (New York: Macmillan Co., 1950), pp. 77-84.

60. Stouffer, Samuel A., and others. Measurement and Prediction (Princeton, N. J.: Princeton University Press, 1950).

61. Thurstone, L. L. "Attitudes Can Be Measured," American Journal of Sociology, XXXIII (1928).

62. Thurstone, L. L. and Chave, E. J. Measurement of Attitude Toward the Church (Chicago: University of Chicago Press, 1929).

63. Trillingham, C. C. Organization and Administration of Curriculum Programs (Los Angeles: University of Southern California, 1934).

64. Walker, Willard. The Sociology of Teaching (New York: John Wiley and Sons, 1932).

65. Wang, Charles K. A. "Suggested Criteria for Writing Attitude Statements," Journal of Social Psychology, III (August 1932), pp. 367-373.

66. Wells, George M. and Green, Lois. "Attitudes Favorable to Change," National Association of Secondary-School Principals Bulletin, XXXIV (December 1950), pp. 141-150.

67. Wickman, E. K. Children's Behavior and Teacher's Attitudes (New York: The Commonwealth Fund, 1928).

APPENDIX

INTERVIEW FORM WITH CATEGORIES FOR CODING INTERVIEW DATA

1. How do you feel that the pupils in your school get along together?
2. How often are you able to help pupils with personal problems?
3. In general, how do you handle pupils when they want to do something you feel is disturbing to the group?
4. How much value do you find in planning with your pupils for class activities?
5. How well do pupils understand and accept what you are trying to do for them?
6. What contact do you have with the pupils in the community?
7. During the school year how much time do you spend outside school with your teacher friends? How much with other friends?
8. How well do you feel that teachers work together in your school?
9. Do you feel that teachers have a real part in determining the policies of the curriculum?
10. How well do the administrators understand and recognize your teaching situation?
11. Have you been encouraged by your administrator in dealing with a school problem?
12. How frequently have you been consulted about curriculum plans?
13. To what extent are you active in community life?
14. Do you have any fairly close friends in the community?
15. Are there some teachers in the school with whom you regularly associate?
16. Whose opinion on school matters would mean the most to you?
17. How do you feel about this curriculum study as compared to other work for class preparation you might be doing?
18. What help do you look for in actual classroom teaching as a result of this study?
19. Do family or home responsibilities make a difference in your share in the study?
20. How free do you feel to make suggestions for curriculum change?
21. If you had your choice, what size group would you like to work with?
22. What suggestions do you have for curriculum revision as you see the needs?

DIETARY PATTERNS OF NEGROES IN THE COMMUNITY OF SCOTLANDVILLE, LOUISIANA*

P. E. THRIFT and R. B. WARD **
Southern University
Baton Rouge, Louisiana

PART I

THE PURPOSE of this study was to determine the food habits of Negroes in Louisiana. It was felt that there was a real need for such a study since no such information was available and the findings should provide basic information for teaching material for the food and nutrition classes taught in Southern University as well as other areas.

Five hundred families, residents of the community of Scotlandville, Louisiana, were chosen for study. This choice was made for two reasons: first, because 95 percent of the total population are Negroes; and second, because of the proximity of the community to Southern University. Scotlandville, the site of Southern University, a state-supported school for Negroes, has a population of 3593 and is located on the Mississippi River six miles north of Baton Rouge, the capital of Louisiana. There are many large industrial plants situated along the river which provide many sources of employment.

The dietary pattern data were collected during the month of June, September, and October of 1949, and June, September, and October of 1950. The two points of time were used to bring into consideration seasonal foods and gardens. The method used for securing the information was a personal interview.

In order to get an accurate record of what the members of the family ate, personal interviews were held with each family by a trained consultant who filled out the questionnaire with the mother or a member of the family who served in the capacity of the mother. Fifty percent of the families in this study were contacted by the writers.

Factors included in this study which affected their meal pattern were income, education, religion, custom, and food dislikes. Therefore, these questions were included in the questionnaire. The persons interviewed were also asked to name the foods they usually ate for a weekday breakfast, lunch, and dinner. In instances where the menu differed on Sundays for breakfast, lunch, and dinner, the foods were listed. They were also asked what special foods they ate on holidays. Other questions asked were the following:

1. What is the weekly and/or monthly income?
2. How much do you spend for food per week and/or per month?
3. How often is each food eaten daily and/or weekly?
4. How are these foods prepared?
5. How often do you serve meatless dinner?
6. What meat substitutes do you use?
7. What foods do you dislike?
8. What foods do you eat between meals?
9. How many meals are eaten daily?
10. Which is the heaviest meal?

The material from this dietary study was tabulated by placing foods into their major groups. Meats included fish and poultry. Vegetables were divided into four categories: cooked green vegetables, other cooked vegetables, raw and canned vegetables, and legumes (beans, red and white). Fruits were fresh, canned, and dried. Dairy products were milk (fresh, canned, and dried), butter, and cheese. Cereal products were bread (white, whole wheat, and corn), prepared and unprepared; rice, macaroni, spaghetti, and hominy grits. The beverages included coffee, tea, and cocoa. Other foods listed were commercial puddings, desserts, molasses, jams, jellies and preserves, which are rich sources of sugar. Eggs and bacon were listed separately.

---

* This study of dietary patterns is being followed by another on the health status of fifty families in the community. The individuals for the health study were chosen from the five hundred families who participated in the meal pattern study.

**The authors acknowledge with appreciation the following: Dr. Elton Harrison, chairman of the Southern University Carnegie Grants-in-Aid Committee, for his efforts in making this study possible; Mrs. Eva Mae Brown Fields and the junior and senior students enrolled in the Division of Home Economics in 1949, for their assitance in collecting the data on the 500 families for Part I; Mrs. Dorothy Moschette, Associate Professor of Home Economics, Louisiana State University, for her professional service; Dr. W. R. Harrison, Professor of Sociology, Southern University, for techniques of research; and in Part II, Dr. R. M. Baranco and Nurse Hope for their technical assistance; Dr. L. R. Posey, Jr., for advice on graphs, and Dr. Walker for his criticism of the manuscript.

Summary of the Data

Of the five hundred families included in this study, the average size of the family was 4.5 persons. The largest family was fourteen in number and the smallest two.

Income

The average per annum income for the five hundred families was $1941.44. The range was from $280 to $9000 per year. Of the five hundred families in the community of Scotlandville, one percent was earning less than five hundred dollars; 12 percent of the families were earning between five hundred and one thousand dollars per annum; 14 percent were receiving between one thousand and fifteen hundred per annum; 21 percent were receiving between fifteen hundred and two thousand dollars per annum; 20 percent were receiving between two thousand and twenty-five hundred; 5 percent were receiving between twenty-five hundred and three thousand; 4 percent were receiving between three thousand and thirty-five hundred; 3 percent were receiving between thirty-five hundred and four thousand; 1 percent were receiving between four thousand and forty-five hundred; 2 percent were receiving between forty-five hundred and five thousand; 1 percent were receiving five thousand and over; 14 percent were unemployed. A report was not obtained from 7 percent of the families. (See Table I.)

It was noted that the families' income, with about 42 percent of the wives employed outside the home, was less than one thousand per annum. This number was 8 percent of the total group. Wives were not employed outside of the home when the husband's income reached the four thousand five hundred bracket.

Education

Eleven percent of the wives and 7 percent of the husbands had completed the eighth grade. Three percent of the wives and 2 percent of the husbands had completed high school. Five percent of the wives and 2 percent of the husbands had completed college.

Religion

Out of the 500 families studied, the Baptist faith dominated for both husbands (58 percent) and wives (77 percent). Methodism ranked second for husbands (10 percent), wives (11 percent). Roman Catholic membership ranked third for husbands (2 percent) and wives (1 percent). Other faiths were represented in order listed: Presbyterian, Holiness, Lutheran, Congregational, and Jehovah Witnesses.

Occupation

The chief occupations reported were the following: unskilled laborers, 51 percent husbands and 6 percent wives; skilled laborers, 10 percent for husbands and 1 percent for wives; professional workers, 6 percent for husbands and 12 percent for wives; unclassified, 1 percent for husbands; 12 percent of the husbands were unemployed. Only 19 percent of the wives were employed outside of the home.

Place of Birth

Sixty-four percent of the husbands and 80 percent of the wives were born in Louisiana. Other states represented were Mississippi, Texas, Ohio, Alabama, Georgia, Arkansas, North Carolina, South Carolina, and Wisconsin.

Typical Meals

Breakfast

The following is a menu of a typical breakfast eaten by five hundred families: eggs, bacon, grits, bread, and coffee. Thirteen percent ate fruit in some form; 32 percent drank milk as a beverage; 2 percent drank cocoa; 14 percent ate cereals other than grits. (See Table II.)

Lunch

Twenty-one percent ate left-overs for lunch; 22 percent ate sandwiches; 22 percent ate knick-knacks; 34 percent did not eat lunch. (See Table III.)

Dinner

The following is a typical weekday dinner menu: lean meat, rice, one green or starch vegetable, and corn bread.

Factors that Influenced the Meal Pattern

Income

The average per annum income for the families was $1941.44. The average amount spent for food was $752.97.

There are wide variations in the amount of

TABLE I

INCOME OF 500 FAMILIES

Range of Income	*Number*	*Percent*
$ 0 - $ 500	5	1%
$ 500 - $1000	62	12%
$1000 - $1500	68	14%
$1500 - $2000	106	21%
$2000 - $2500	100	20%
$2500 - $3000	27	5%
$3000 - $3500	21	4%
$3500 - $4000	17	3%
$4000 - $4500	2	1%
$4500 - $5000	9	2%
$5000 - and above	15	3%
Number unemployed	68	14%

TABLE II

FREQUENCY OF FOODS EATEN FOR BREAKFAST

Foods	Number	Percent
Bread	426	85%
Grits	395	79%
Eggs	333	67%
Bacon	256	51%
Coffee	206	41%
Butter	205	41%
Milk	160	32%
Cereals	72	14%
Fruit	63	13%
Juice	35	7%
Syrup, preserves, etc.	11	2%
Cocoa	8	2%

TABLE III

FREQUENCY OF FOODS CHOSEN FOR LUNCH BY THE WIVES AND CHILDREN UNDER SCHOOL AGE*

Foods	Number	Percent
Sandwiches	110	22%
Left-overs	109	22%
Cold drinks	33	7%
Fruit	16	3%
Milk	15	3%
Eggs	13	2%
Bread	9	1%
Jam	6	1%
Crackers	5	1%
Coffee	4	1%
Juices	4	1%
Cereals	4	1%
Salad	2	1%
Butter	2	1%

*Ninety-five percent of the husbands and the children of school age ate their lunches outside of the home.

money spent for food for families of the same size, with attendant differences in expenditures for particular foods. These differences are minimized and tend to disappear with marked downward shifts in food costs, just as they are accentuated with marked upward trends in cost of certain foods. In families where the income is $2500 or more, as much as three to four dollars is spent for fruit. Fruits out of season are served for breakfast and often for the midday meal. Expensive meats such as roasts and steaks are on the daily dinner menu. With the small food budget for the very low income level, the choice is limited; red beans cooked with salt pork and served with rice constitute the substantial part of the meal.

A large number of the families, eighty-eight percent, plan their meals by the day; 6 percent plan their meals by the week; 6 percent did not plan meals at all. Sixty-seven percent shopped daily; 6 percent shopped three times per week; 8 percent twice per week; 19 percent shopped once per week. Inadequate refrigeration is one factor that was responsible for frequent shopping. These families shop every day during the week, including Sunday. However, Saturday is the most popular day. Eighty percent of them market in the morning; 20 percent marketed in the evening; 74 percent market at the corner grocery store; 10 percent marketed at the chain store; and 15 percent marketed at the chain and corner stores. The location and method of transportation may prevent a large percent from marketing at the chain stores.

Very little variety was used in the diet. The study revealed that a majority of the families had not seen nor heard of such foods as cauliflower, broccoli, eggplant, and brussel sprouts. Those foods that were known were usually prepared the same way.

Education

Education usually has some affect upon the dietary practices of individuals; however, this study showed that although there were families who knew what constituted a well-balanced diet, they actually ate no better than those who knew little about balanced meals.

Religion

Often religion prescribes or prohibits the eating of certain foods. However, in this study religion had little or no important bearing on these families' dietary habits or practices.

Custom

The type of meal itself, as well as the characteristic dishes included, is often largely determined by family custom. On Sunday, ham and eggs are included in the breakfast menu. Other Sunday dishes are green peas, potato salad, and chicken. In 85 percent of the homes corn bread is served with the dinner meal and grits are served daily for breakfast.

Food Dislikes

Likes and dislikes often exert undue influence, and many people make personal likes and dislikes the basis of the inclusion or exclusion of a food in their diet. The failure to include cereals, especially whole grain cereals, brown bread, cheese, milk, tea, lettuce, carrots, asparagus, broccoli, and eggplant revealed this fact. (See Table IV.)

Conclusion

The data collected from the 500 families showed that they ate the following foods for breakfast: eggs, bacon, grits, bread, and coffee. Only a very small percentage ate fruit in any form for breakfast. Fourteen percent of the families ate cereals other than grits. Ninety-nine percent ate hominy grits, not as a breakfast cereal, but as a part of the main meal. Cornflakes ranked highest in preference but whole grain cereals were excluded almost entirely. The beverages included in the breakfast meal, in the order of frequency were coffee, milk, and cocoa. The study also revealed that one-fourth of the families did not include a beverage.

The Sunday breakfast was very similar to the weekday meal pattern, with the exception that ham and toast were also served.

A definite luncheon pattern did not exist among the families; however, families who had lunches ate left-overs, sandwiches, and knickknacks. The fact that 95% of the husbands were employed away from home and carried packed lunches or ate on their jobs seemed to be the factor responsible for this.

The dinner menu always included lean meats, fish, or fowl. The lean meats used in order of their preference were pork, beef, and lamb. Ninety-eight percent of the families invariably used rice, served as a vegetable. Red beans and sweet potatoes were frequently served, often in the same meal. White potatoes were served, but less frequently when sweet potatoes were in season. The green vegetables used for dinner were turnips, mustards, or collard greens. These vegetables were always cooked with salt pork.

The Sunday dinner was more elaborate than the weekday dinner. A typical Sunday dinner consisted of chicken, green peas, rice, potato salad without the lettuce, hot bread, and dessert. Canned peaches and/or pudding appeared

TABLE IV

FOOD DISLIKES

Foods	Number	Percent
Meats:		
Mutton	227	45%
Lamb (roasts, chops, steaks)	219	44%
Veal (roasts, chops, steaks)	37	7%
Hamburger	24	5%
Pork (roasts, chops, steaks, ribs)	19	4%
Liver	15	3%
Ham	14	3%
Beef (roasts, chops, steaks)	10	2%
Sausage	5	1%
Bacon	3	1%
Chicken	2	1%
Seafoods:		
Oysters	115	23%
Crab	43	9%
Shrimp	37	7%
Fish	5	1%
Beverages:		
Tea	31	6%
Coffee	19	4%
Chocolate	19	4%
Cocoa	17	3%
Milk	7	1%
Vegetables:		
Broccoli	345	69%
Cauliflower	342	68%
Asparagus	333	67%
Eggplant	252	54%
Carrots	188	38%
Beets	82	16%
Celery	66	13%
Cabbage	59	12%
String beans	18	4%
White beans	13	3%
Okra	13	3%
Macaroni	12	2%
Sweet potatoes	11	2%
Lettuce	10	2%
Grits	9	2%

TABLE IV (Continued)

Foods	Number	Percent
Spaghetti	9	2%
Corn	9	2%
Onions	7	1%
Rice	7	1%
White potatoes	5	1%
Eggs	12	2%
Cheese	9	2%
Cereal and Cereal Products:		
Petti Johns	214	43%
All Bran	212	42%
Wheatina	203	41%
Farina	160	32%
Cream of Wheat	154	31%
Kix	138	28%
Grape Nut Flakes	111	22%
Wheaties	108	22%
Oat Meal	83	17%
Corn Flakes	38	8%
Other cereals	30	6%
Brown Bread	150	30%

TABLE V

THE AVERAGE AMOUNT OF MONEY SPENT FOR FOOD PER YEAR ACCORDING TO INCOME

Income per Year	Percent of Families	Ave. Amount Spent for Food per Year	Percent Spent for Food
$ 0 - $ 500	1%	$ 402.80	86%
$ 500 - $1000	12%	544.22	54%
$1000 - $1500	14%	671.32	45%
$1500 - $2000	21%	775.96	39%
$2000 - $2500	20%	940.46	38%
$2500 - $3000	5%	986.66	33%
$3000 - $3500	4%	970.49	28%
$3500 - $4000	3%	732.70	18%
$4000 - $4500	1%	1200.00	27%
$4500 - $5000	2%	1074.66	21%
$5000 and Above	3%	1062.20	21%

to be the favorite dessert.

This study revealed that the characteristic meal pattern for the families studied relative to the three daily meals was a heavy breakfast, a light lunch, and a heavy dinner.

Food dislikes caused the exclusion of many essential foods from the family diet. This fact was revealed by the failure of many families to include cereals, brown bread, cheese, milk, lettuce, carrots, asparagus, broccoli, and eggplant.

In one percent of the families it was revealed that after taking care of the food budget only 14% of the total income was left for clothing, shelter, and other essentials. Families with an income of $2000 to $2500 per year spent less than 50% of the total income for food. In families with an income of $3000 to $5000 and above per year only 5 percent spent as much as 30% for food. (See Table V.)

A STUDY OF THE RELATIONSHIP BETWEEN THE DIETS AND THE HEALTH OF 25 NEGRO FAMILIES IN SCOTLANDVILLE, LOUISIANA

P. E. THRIFT and R. B. WARD
Southern University
Baton Rouge, Louisiana

PART II

Purpose of Study

ORIGINALLY, the purpose of this study was to determine the health status of fifty families in the community of fifty families in the community of Scotlandville, Louisiana. This was to be accomplished by evaluating the diets of these families and by subjecting them to a physical examination.

Efforts were made to study 50 families who participated in Part I of the dietary pattern study, but only twenty-five families ultimately agreed to cooperate in this investigation.

Procedure

The data were collected during the month of July, 1951. In order to ascertain what the 25 families ate, a complete dietary record was carefully kept by the head of each family for a period of one week. Each family filled out the dietary record by checking the form daily. The dietary record sheet for each day included the name of each meal, a list of foods eaten, the quantity of the food, and in addition, the kind and amount eaten between meals. Space was provided on the record sheet for additional foods if needed. To insure accuracy, consultants checked daily with the families on the filling out of their records. At the end of the seven day period, dietary records were collected and evaluated.

Accepted standards were followed.*

The fifty families were subjected to a medical history which included a physical examination given by a competent physician.

Findings

Not only was the percentage of the group meeting caloric requirements too low as shown in Chart 1, but the group fell far short on other required nutrients. The study also revealed a lack of balance between carbohydrates, fats and proteins in every family studied. Fats were used excessively in 100% of the cases. This can be accounted for by the fact that the method most frequently used for preparing meats was frying and that all vegetables used were cooked with fat pork.

The dietaries studied showed generally that the amount of vitamin A consumed by these families came chiefly from turnip or mustard greens, sweet potatoes, and cushaw. Ascorbic acid was derived from cabbage, grapefruit, orange juice, and raw tomatoes. Thiamin, riboflavin, and niacin came from lean meats including chicken and liver; enriched breads.

Although 20% of the families were not receiving adequate amounts of proteins, the proteins consumed were complete. They came chiefly from lean meats, fish, and poultry. Dried peas and beans were included almost daily in the diets of 64% of the families.

As seen in Chart 1, 60% of the families were consuming less than the recommended amounts of calcium. This was evident from the small amount of milk used in the diet. Families receiving adequate amounts of calcium obtained it mostly from evaporated milk, mustard and/or turnip greens. Milk, green peas, sweet corn, eggs, and canned fish helped supply sufficient amounts of phosphorus for 60% of the families. The variety of leafy greens, enriched breads, and liver provided the chief sources of iron for 80% of the families.

When a family's food consumption record indicates that it met the recommendations in nine or more of the nutrients as shown in Chart 2, it was rated good. It was rated fair if it met from five to eight of the nutrients; poor if it met the recommendations in less than five nutrients.

Summary

Sixty-eight percent of the families were consuming less calories than recommended and 64% ate less carbohydrates than was necessary. The study also revealed that their meals lacked proper balance.

Twenty percent of the families were not re-

---

*Donelson, Eva G. and Leichsenring, Jane M. "Food Composition Table for Short Methods of Dietary Analysis," reprint from Journal of the American Dietetic Association, XXI (July-August 1945).

CHART 1

PERCENT OF FAMILIES MEETING INTAKE RECOMMENDATIONS

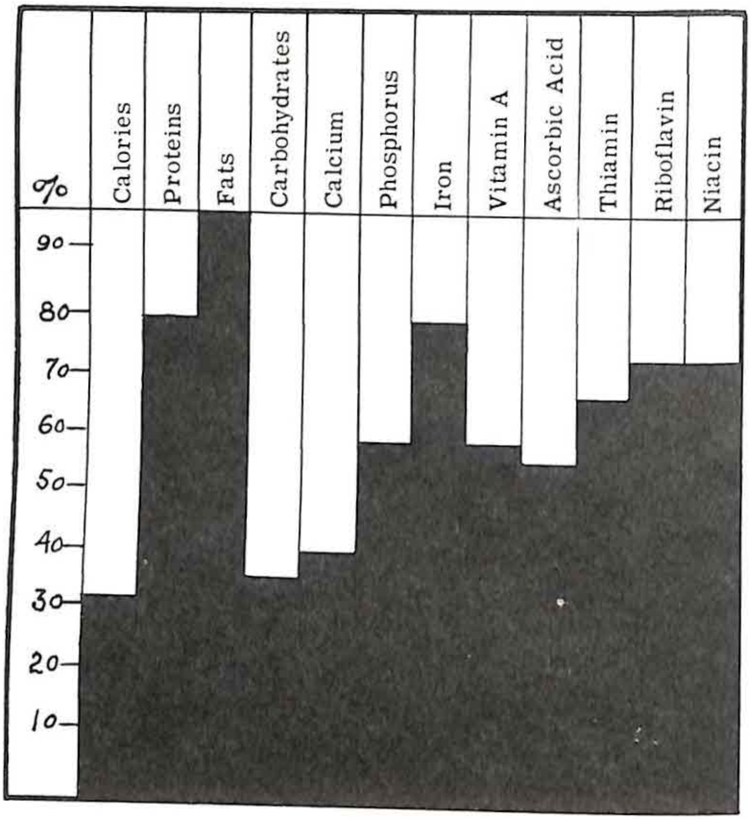
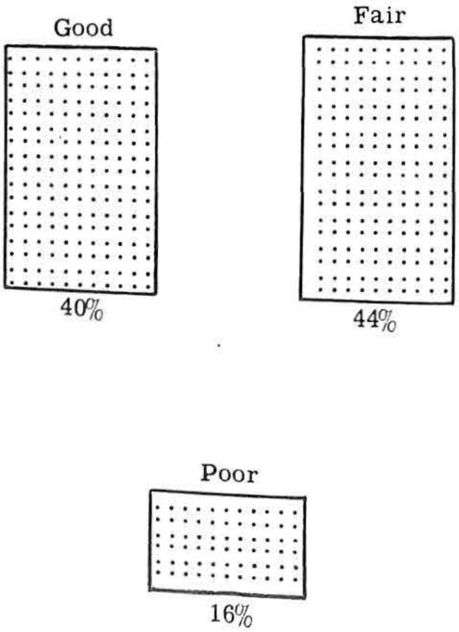


CHART 2

EVALUATION OF DIETARIES



ceiving adequate amounts of proteins. However, those received were complete proteins since they came chiefly from lean meats, fish, and poultry. Dried beans and peas were included almost daily in the diets of 64% of the families.

Sixty percent of the families were consuming less than the recommended amounts of calcium. This was evident because of the small amounts of milk used. Forty percent were not receiving the recommended amounts of phosphorus. In iron 20% were not reaching the recommended intake.

Forty percent were receiving less than the recommended amounts of vitamin A; 44% were receiving less than the recommended amounts of ascorbic acid (vitamin C); 32% receiving less than the recommended amounts of thiamin (vitamin B_1); 28% receiving less than the recommended amounts of riboflavin (vitamin B_2); and 28% were receiving less than the recommended amounts of niacin. The chief source was enriched bread; in some instances 4 to 6 slices were consumed at a single meal. Other sources were leafy greens and lean meats.

It was interesting to note that no family excluded breakfast. Only two meals were served on Sundays; this included a heavy breakfast and heavy dinner, lunch being omitted.

The physical examination for the twenty-five persons studied showed the following results: 20% were underweight, 52% were overweight, and 52% had dental caries.

There was only one case of diabetes. The other subjects studied had no specific diseases. Ninety percent, however, did have ailments now known definitely to have dietary origin, these being fatigue, frequent headaches, and constipation.

REVERSALS, REVERSALS, REVERSALS!

HELEN KENNEDY
Los Angeles State College

OVER A PERIOD of many years, particularly during the last twenty or twenty-five, the educational and psychological literature has contained numerous articles on strephosymbolia—twisted symbols. During this time little advance has been made in the understanding of the condition. It may well be that the lack of advancement is due to the fact that little effort has been made to differentiate the types of reversals and to ascertain the frequency of their occurrence among a group of normal children of given age levels. Such a study was made in 1934 but the data were never reported in the literature. Although these data are old they are being presented at this time in the hope that persons continuing research in this area may find them useful.

The first step of such a study was, of course, to devise a test instrument with which various reversal types could be differentiated and the frequency of their occurrence could be studied with children of different age levels. In Part One of the test, twelve exercises were presented, following three practice exercises. In each instance the key symbol was presented as the left column. To the right, separated by a double line, were twelve symbols like the key symbol except that only three of the twelve were in the same orientation. Of the other nine symbols, three were inverted, three were reversed, and three were inverted-reversed. The twelve symbols used as key stimuli included five letters, two numbers and five designs. The five letters were “p”, “u”, “f”, “d” and “t”. These particular letters were selected for the obvious reason of the confusions which occur with other letters. Both “p” and “d” are almost exact orientational variations of each other and of “b” and “q”; “f” and “t” are similar except for orientational variations and are also similar to “j”. The “u” and “n” are inverted-reversed images as are “9” and “6”. The other stimulus number was “5” which is very like “2” when inverted.

Part Two consisted of twenty-four exercises, preceded by three practice exercises. In the left hand column the stimulus symbol, this time a common word, was presented. The words varied in length from two to five letters. To the right of a double line were six forms from which the child was to select the one like the stimulus pattern. In each exercise there was the correct pattern, an irrelevant pattern and one each of four major reversal types.

The irrelevant stimulus pattern which was included in each of the exercises in Parts Two and Three was a word pattern of the same general configuration but which did not contain any of the same elements. It was assumed that if no discrimination was taking place this pattern would receive a comparable share of responses.

The first type of reversal was called a letter change since the disorientation involved only one letter in the word. The change was either a reversal, an inversion, or an inverted-reversal. Examples of these changes are, respectively, “big-dig”, “big-pig” and “dig-pig”. The stimulus words were selected so that the letter with which the change occurred might be in the initial, the internal or the terminal position in the word. The second type was called the reversed form. In this situation each letter remained correctly oriented but the sequence of the letters was reversed, i. e., “was-saw”. The third type was called the mirrored image. Here again the orientational change occurred either by inversion, reversal or inverted-reversal. These effects may be obtained by the use of mirrors. To observe the inverted mirrored image the mirror should be placed above or below the symbol. The reversed mirrored image may be seen by placing the mirror either to the right, the left or parallel to the symbol. The simplest means of seeing the inverted-reversed form is to turn the page upside down. The last major reversal type was termed transposition. Eight types of transpositions were identified and included in the final form of the test. The transposition types identified were as follows: 1) initial letter to an internal position; 2) initial letter to terminal position; 3) terminal letter to initial position; 4) terminal letter to internal position; 5) internal letter to a different internal position; 6) internal letter to initial position; 7) internal letter to terminal position; and 8) miscellaneous changes involving several of the others or more than one letter. In the data to be reported in this paper the types described above as 6, 7, and 8 were all included under miscellaneous.

With two letter words it is, of course, impossible to have a transposition other than a reversed form. Therefore, in the first six exercises of Part Two, there are two mirrored forms rather than two reversed forms of the word.

Part Three of the test is based on the same

TABLE I

TOTAL NUMBER OF TEST PAPERS AVAILABLE FOR USE AFTER MAKING EXCLUSIONS BECAUSE OF UNSUITABILITY FOR COMPUTATION OF AVERAGES, BY GRADE LEVELS

Grade Level	Total No. of Tests Administered		Test Papers Retained					
			Part I		Part II		Part III	
	Form 1	Form 2	Form 1	Form 2	Form 1	Form 2	Form 1	Form 2
LK (Low Kindergarten)	29	45	26	38	16	26	*	*
HK (High Kindergarten)	45	43	36	36	25	28	*	*
L1 (Low First)	36	37	35	35	32	31	*	*
H1 (High First)	64	69	63	67	57	67	24	32
L2 (Low Second)	38	48	38	48	38	48	22	26
H2 (High Second)	48	96	48	96	48	96	39	78
	260	338	246	320	216	296	85	136

*Note: Only a short form of the test, i.e., parts I and II, were given to the lowest grade levels, as part III requires more word knowledge than is expected of children in those grades.

TABLE II

PERCENTAGE OF TOTAL POSSIBLE RESPONSES IN EACH CATEGORY WHICH IS FOUND WITH THE AVERAGE OF EACH CLASS RESPONSE TO THAT CATEGORY, IN PART ONE

	Correct Responses		Incorrect Responses		Inverted Responses		Reversed Responses		Inverted-Reversed Responses	
	Form 1	Form 2	Form 1	Form 2	Form 1	Form 2	Form 1	Form 2	Form 1	Form 2
LK	63.8	65.8	22.1	15.2	20.2	12.6	31.4	28.4	13.8	5.8
HK	64.0	67.4	20.2	20.0	17.0	13.2	35.6	38.8	8.2	8.4
L1	83.4	77.6	12.8	8.9	10.5	10.9	23.0	13.3	4.8	2.9
H1	79.2	80.4	10.5	8.2	9.7	9.1	16.6	13.2	5.3	2.5
L2	88.7	91.8	2.8	5.1	4.7	9.3	3.5	5.9	.15-	.27
H2	95.0	90.2	3.07	6.4	5.7	10.9-	3.3	7.0	.29	1.5-

TABLE III

PERCENTAGE OF TOTAL POSSIBLE RESPONSES IN EACH CATEGORY WHICH IS FOUND WITH THE AVERAGE OF EACH CLASS RESPONSES TO THAT CATEGORY

	Correct Responses		Incorrect Responses		Irrelevant Responses		Omissions	
	Form 1	Form 2	Form 1	Form 2	Form 1	Form 2	Form 1	Form 2
				PART TWO				
LK	38.4	40.3	20.2	21.1	1.8	2.1	9.9	14.6
HK	42.0	50.7	16.1	19.4	2.3	1.8	4.7	10.7
L1	66.2	75.0	13.3	8.8	.25	.4	14.2	9.3
H1	77.5	84.6	8.7	7.4	.22	.2	3.8	2.7
L2	92.8	89.6	3.4	3.7	00	.1	2.5	2.3
H2	93.0	92.4	3.4	2.6	00	.04	1.9	2.9
				PART THREE				
H1	24.7	26.6	15.0	14.4	7.2	7.8	16.3	13.3
L2	38.7	32.3	9.5	9.8-	6.1	4.8	9.7	3.7
H2	56.0	53.8	7.3	9.3-	.86	2.4	9.6	6.1

master plan as Part Two. It contains five more exercises than does Part Two and the words vary in length from three to nine letters. The stimulus pattern in Part Three is a picture and the child is requested to select the word which represents the picture. All words for the test are taken from Gates Primary Word List.

Two forms of the test were developed from the same master plan. Some of the same words were used in the two forms which made it possible to check the effect of position of the word in the exercise and its relative position to other types of errors. Naturally the administration of the test was also facilitated by having the two forms.

It will be recalled that one of the purposes of the study was to determine the relative frequency of occurrence of the various types of reversals among children of given school ages. For this purpose, permission was obtained to test all of the kindergarten, first and second grade children in three of the Pasadena City Schools. Not all papers which were obtained were found to be usable. It was possible, however, to use the responses in one part of the test even though another part of the test had to be disregarded. Papers were considered as acceptable if at least half of the exercises of a part had been attempted, if not more than one third of the irrelevant responses were marked and if there was no evidence of a patterning of the responses according to position on the paper, i. e., always the first symbol. After discarding all papers which did not meet the above criteria, there remained for study, in computing the class group averages, the numbers of test papers for the three parts of the test as reported in Table I.

Incorrect responses made by the children are of little significance unless the qualitative and quantitative nature of the errors may be determined. Hence, it is necessary to consider the errors in their various type classifications and sub-classifications. Since, however, the tests did not present the same opportunities for making the several types of errors it is desirable to report all data in percentages. Such will be the procedure in Tables II through VI. The base will be the total opportunities offered for responses of the type under consideration and the reported percentage is the proportion that the responses are of that total base.

For example, in Part One of the test there are 36 possible correct responses. If the average of the group is 22.97 correct responses, then the tendency to respond correctly is measured by determining the percentage which 22.97 is of 36, viz: 73.8%.

Table II is concerned with the results from Part One of the test. In the first two columns are reported the relationship between the average number of correct responses, for each form of the test, for each grade group and the total number of opportunities for such responses. The next two columns, in like manner, indicate that percentage which the average number of incorrect responses is of the total number of incorrect responses which it is possible for one to make. The remaining columns concern the subdivisions of the incorrect responses. The data in them were calculated in the same manner as in the first four columns.

In Part One of the test there have not been included any possibilities for an irrelevant response. Furthermore, in the scoring of these exercises no consideration was made of a failure to mark all or part of an exercise, as Part One of the test was scored exclusively for responses actually made. Consideration of irrelevant responses and omissions, however, is given in scoring Parts Two and Three.

Table III indicates the tendency at each grade level to make correct, incorrect or irrelevant responses or to omit making any response with Parts Two and Three of both forms of the test. The data reported in this table were derived by a method similar to that described for the derivation of the data in Table II. Sub-classifications of the incorrect responses will be presented in Tables IV, V and VI.

In Table IV is presented the percentage of frequency of occurrence, in relation to the opportunities for each, of the four major reversal types. Tables V and VI present the breakdown of these data into the subclassification previously identified.

Although it is recognized that the number of subjects involved in this study was not adequate to permit the forming of any definite conclusions, certain indications are more than suggestive and worthy of further exploration. Two questions seem pertinent with regard to the maturation process which ultimately results in the almost complete disappearance of the reversal tendency in the work of the average or superior child. One of the questions considers whether the frequency of occurrence of the reversal situation as a whole significantly and steadily decreases as the children become physically and educationally more mature. A second question is whether any particular reversal type deviates significantly from the general pattern of the whole reversal situation.

The first question may be answered emphatically in the affirmative. Consideration of the data presented in the preceding tables indicates that the type of discrimination required for success with this test seems to have had its origin at some time previous to the child's entrance to the kindergarten. It seems to be indicated that correct discrimination was not to any great extent chance occurrence. The correct responses were selected with much greater frequency

TABLE IV

PERCENTAGE OF TOTAL POSSIBLE RESPONSES IN EACH CATEGORY WHICH IS FOUND WITH THE AVERAGE OF EACH CLASS RESPONSES TO THAT CATEGORY

	Letter Change Responses		Reversed Form Responses		Mirrored Responses		Transposed Responses	
	Form 1	Form 2	Form 1	Form 2	Form 1	Form 2	Form 1	Form 2
				PART TWO				
LK	30.2	25.5	25.5	29.4	13.7	13.9	19.4	22.0
HK	29.7	27.4	18.3	23.6	10.6	15.2	14.7	18.8
L1	25.2	21.2	12.0	6.0	6.4	2.3	9.0	8.2
H1	24.0	20.3	5.8	5.9	4.7	3.1	4.8	4.8
L2	13.3	12.2	.66	2.2	.68	1.04	1.4	3.4
H2	10.7	7.6	1.39	2.0	.74	.9	2.6	2.9
				PART THREE				
H1	20.4	17.3	10.5	10.1	8.5	9.5	13.5	13.5-
L2	14.2	16.0	4.4	4.8	5.7	4.4	7.8	14.6
H2	12.3	17.7	2.2	3.3	1.9	2.5	9.0	13.5

TABLE V

PERCENTAGE OF TOTAL POSSIBLE RESPONSES IN EACH CATEGORY WHICH IS FOUND WITH THE AVERAGE OF EACH CLASS RESPONSES TO THAT CATEGORY

	Letter Changes						Mirrored Forms					
	Inverted		Reversed		Inv. -Rev.		Inverted		Reversed		Inv. -Rev.	
	Forms		Forms		Forms		Forms		Forms		Forms	
	1	2	1	2	1	2	1	2	1	2	1	2
	PART TWO											
LK	27.4	18.3	36.0	43.4	27.4	18.0	9.4	11.1	23.7	23.4	9.3	8.0
HK	25.5	18.8	39.0	47.0	24.5	20.7	9.5	8.5	17.2	28.9	6.7	8.3
L1	21.9	17.8	33.2	35.3	20.4	13.6	7.0	4.9	9.7	2.6	3.6	.3
H1	20.0	14.4	33.8	37.6	18.2	12.6	5.8	3.4	7.5	4.8	2.1	1.6
L2	10.5	8.4	21.0	26.8	8.2	4.4	00	1.8	1.8	1.3	.2	.4
H2	7.5	5.5	17.1	15.3	7.3	3.5	1.1	.8	1.0	1.8	.4	.3
	PART THREE											
H1	17.9	16.3	26.0	20.7	18.5	14.2	7.9	8.8	10.8	11.6	6.1	8.0
L2	12.3	11.1	15.9	22.9	14.4	12.5	5.8	4.0	8.2	5.2	2.5	4.0
H2	11.0	17.8	14.6	23.1	11.9	10.8	3.7	2.8	1.3	2.8	.4	2.0

TABLE VI

PERCENTAGE OF TOTAL POSSIBLE RESPONSES IN EACH CATEGORY WHICH IS FOUND WITH THE AVERAGE OF EACH CLASS RESPONSES TO THAT CATEGORY

	Initial to Internal Position		Initial to Terminal Position		Terminal to Initial Position		Terminal to Internal Position		Internal to Internal Position		Miscel- laneous Position	
	Forms		Forms		Forms		Forms		Forms		Forms	
	1	2	1	2	1	2	1	2	1	2	1	2
PART TWO												
LK	20.2	30.4	6.2	24.3	8.3	17.3	37.3	25.7	37.5	38.5	16.2	8.5
HK	8.0	19.0	6.0	17.9	5.3	19.6	26.6	23.8	28.0	26.8	19.2	12.8
L1	6.2	6.5	3.1	4.3	5.2	3.2	17.7	15.0	33.4	19.5	5.6	5.2
H1	2.6	6.0	00	00	1.2	00	14.0	8.0	12.3	12.7	3.5	3.9
L2	.7	1.4	00	.7	.9	00	4.4	4.9	2.6	13.5	1.1	2.5
H2	.5	2.0	00	00	00	00	4.9	5.0	16.7	11.5	2.5	1.6
PART THREE												
H1	7.5	2.5	13.5	13.3	5.6	6.2	24.2	34.4	13.9	14.4	13.4	9.4
L2	2.7	1.6	5.7	8.0	4.5	4.0	13.6	39.2	13.6	24.0	7.2	8.7
H2	1.5	1.3	6.3	3.5	00	1.6	16.7	39.6	20.0	28.2	9.4	4.5

than the incorrect; and perhaps even more significant was the very low incidence of irrelevant responses. In all three parts of the test there is a steady increase in the percentage of correct responses and equally steady decrease in the incorrect responses.

Further consideration of these data might seem to suggest a problem deserving of serious investigation. There seems to be little difference in the achievement of the two kindergarten levels. Could it be that the kindergarten situation is less productive of growth in the type of discrimination needed for the proper identification of printed word symbols than is the environment from which the child was taken? This situation coupled with the rapid growth which seems to have occurred in the early part of the first grade gives rise to the question of whether a closer coordination between the activities of these grade levels might not be productive of greater growth.

With regard to the second question, whether any particular reversal type deviates significantly from the general pattern of the whole reversal situation, the following observations seem justified. With Part One of the test, it seems that of the three possible error types occurring with equal frequency in the test stimulus, pupils at the lower grade levels chose the reversed form with greatest frequency. The inverted reversed form does not occur with as great frequency and disappears more rapidly than do the other two. The inverted form seems to be more resistant than do the others. All of these errors disappear quite rapidly.

Parts Two and Three of the test offer opportunities to make error of four major types, viz: letter orientation changes, reversed form of the word, mirrored forms of the word and letter transpositions of the word. A significant feature of the data regarding these types of reversals is that they do not occur at any of the grade levels with a frequency which one might have expected from the discussion in the literature regarding the problem. Another important point is that the mirrored form of reversal is indicated to be less common than is any other type. The reversed orientation of letters is the most common and it disappears most slowly. Three of these four major reversal types are subject to greater refinement. Summarization regarding the findings with the sub-classifications follow.

Greater difficulty appears among the letter orientation changes than where the same changes occur with a single symbol stimulus. Among the three sub-types of errors, the reversed form causes the greatest difficulty in both instances. The persistence of this error type is much greater, however, in the case of letter orientation changes.

There is, surprisingly, a small number of errors with the mirrored form of the word. It seems that the gross orientation changes are more easily recognized than those of the same type which involve only a letter of the word. Again the reversed form produces a higher degree of error than the others but rapidly decreases to a comparable level. These data do not appear to indicate that the mirrored forms of the word constitute any very great potential difficulty. There exists, however, a very striking similarity in the contour of the curves of the corresponding sub-types of the mirrored forms, the letter orientation changes and the single symbol type of Part One.

There is rather rapid disappearance of the tendency to transpose letters where the initial letter is concerned. In instances where the transposition involves only the internal letters or an internal and terminal letter, there is a greater tendency for the error type to persist.

One might conclude from these data that the occurrence of reversal errors of all the types studied are perfectly normal in the earliest grades and that it is equally normal for them practically to disappear as the child progresses through the first several grades.

Even though it is granted that some tendency to make errors of the types described is normal at given age levels, the question still remains as to what can be done when the tendency persists. The question might also be raised as to whether the tendency needs to exist to the degree that it does even in the very early grades. Many of the conditions studied have been as poorly defined and as incompletely considered as have the types of reversals, consequently, the findings are inadequate and inconsistent. Perhaps the simplest explanation, and the one which is likely responsible for most of the reversals, seems to have been completely overlooked, at least as far as reports in the literature are concerned. The idea was brought very forcibly to the attention of the writer when a junior high school boy asked whether it made any difference which way the stem and the bulb pointed. A boy is a boy, regardless of whether he is facing north or south, standing on his feet or his head. So it is with the material things with which a child is familiar. Yet in writing, and particularly with printed symbols, we take one symbol, turn it in four directions and call it by four different names. How often is it explained that in the reading of word symbols all the streets are one way, from left to right, at least in the languages to which most of us are exposed.

C Certainly it is not the contention here that so simple an explanation is the answer in all instances. It is well worth a try and will eliminate, for further consideration, many youngsters now classified as having "reversal difficulties." When such children have been removed from consideration by this simple expedient, then those whose problem is more complex may be studied with some hope of determining causative or at least relational factors.

THE IMAGE OF THE INSTRUCTOR AS IT IS RELATED TO CLASS SIZE

JOHN B. HOLLAND
Michigan State College
East Lansing, Michigan

AT ONE TIME the field of educational research included many who, for one reason or another, sought answers to the question of the relationship of class size to various educational factors. This research, which began with the pioneer work of Rice in 1895, includes more than 70 separate studies.[1] These studies have ranged from small and sometimes impressionistic treatments to elaborate and carefully controlled experimental designs. Most of this research has been focused upon the effect of class size on pupil attainment in subject matter. But despite the many notable researches that have been undertaken, the answers are confused and uncertain. It is not surprising that investigation into this important area has been all but abandoned for the last 15 years.

The research that is to be reported herein is a variation upon this old, familiar, but now almost unsung theme. In the beginning, the writer and his colleagues[2] discussed and planned a rather elaborate experimental design with the purpose of answering a number of questions about relationships which might be hypothesized to exist between class size and such factors as pupil attainment, attitudes evoked by the subject matter, evaluation of the class and class procedures, and image of the instructor. For many reasons, including time and availability of research funds and facilities, the total data collected within this design were not analyzed. Inspection of the data, however, indicated that many of the results appeared to add little to the results of other class size studies. In this connection, Otto and Borgersrode have written, "As long as rather conventional, routinized group procedures are used in teaching it is not likely that research will show whether large or small classes are to be preferred."[3] The difficulties of controlling for all other possible factors, together with the fact that the usual class size experiment may not be testing the crucial variables, help to increase the confusion.

One datum, however, appeared to indicate, even by inspection, that certain important and unique differences might exist. This factor was what may be termed the image of the instructor manifested by the student. That is, there appeared to be, and the experimenters themselves also reported, that there was variation in the way that students of different class sections perceived the instructor. These differences, therefore, were investigated more carefully and the results of this analysis furnish the basis for this report.

The Design of the Research

It is not important that the total research plan be reported, since only a segment of it is discussed here. The experimental procedure, however, should be presented.

The original sample for the total research design consisted of some 784 regularly enrolled students in a general education college course called Effective Living.[4] These students were taking the second term of this three term course. Hence, some of them had had previous contact with the experimenters. However, since these students were randomly located in the experimental sections and were drawn, through the chances of registration, out of a total of 2043 students who were enrolled in the second term of the course, the probability of this prior contact was low. Although these students may not be spoken of as constituting a sample of the student body, they were representative, for the most part, of freshmen and sophomore students at Michigan State College. This is true because they were in a required educational program which, although allowing some choice on the part of the student, regularly included more than seventy-five percent of the student body. That is, most students at Michigan State College take this particular course early in their college career. There is no reason to believe that the

---

1. Walter S. Monroe (Ed.), *Encyclopedia of Educational Research* (New York: Macmillan Co., 1950), pp. 212-215.
2. Professors E. B. Blackman, R. G. Davis, R. F. Fink, H. Rosencranz, and Leo A. Haak, Head of the Department of Effective Living, all participated actively in the experiment. The writer is indebted to them for their many suggestions and assistance in planning and carrying out this research.
3. Monroe, *op. cit.*, p. 215.
4. One of seven basic general education courses at Michigan State College, until 1953.

sample used in this experiment differed in any way significant for the research problem from the usual population of freshmen and sophomores in earlier or succeeding years. At the time of the experiment the veteran population of the college was low.

The research plan called for four different kinds of teaching sections and the use of four different instructors as experimenters. The teaching sections were selected to approximate the following kinds: (1) a large lecture-discussion section of approximately 100 students; (2) a medium-size lecture-discussion section of approximately 50 students; (3) a small lecture-discussion section of approximately 25 students; and (4) a small section of 25 students wherein the experimenter was allowed complete freedom to use whatever methods of instruction he desired. The first three types of sections were designed to control, as adequately as possible, the method of instruction. That is, in all these sections the experimenter was instructed to use the lecture method and allow for such discussion as did not prevent his covering the regularly assigned materials.

It is obvious, of course, that the method of instruction could not be held as an absolute constant. That is, a section of 100 students allows, by the very fact of size, less opportunity for a large number of students to participate through discussion or questions. Likewise, the small section of 25 students imposes a less formal atmosphere and allows, in terms of simple economics, greater participation by more students. It should be emphasized here, however, that the experimenters attempted to hold constant, insofar as possible, their own presentations. While they lectured more in the largest sections, they also attempted to create an atmosphere of sufficient permissiveness that discussion could and did take place. In the smaller sections the instructor controlled the discussion in such a way that he could and did repress it when it was necessary to complete the same materials presented in the larger lecture section. Thus, the larger section was probably more permissive than many such sections wherein the instructor lectures without interruption. The smaller sections were probably less permissive than those wherein the instructor attempts to draw out the students and encourages, without limit, their participation in the classroom.

The small “free” experimental section is more difficult to describe. Data were obtained from each experimenter each week on the procedures used in these sections. There was wide variation among instructors. The procedures ranged from full classroom discussion to role-playing, projects, small group work, and greater use of audio-visual aids than the fixed and required materials presented in the other sections. As such they contrasted sharply with the conventional teaching sections used in the experiment.

The instrument used to measure the student's perception or “image” of the instructor consisted of 10/five-point scales. These scales covered a number of aspects of instructor personality and were arranged so that the extreme end-points were identified but the intervening steps had no descriptive designations. For example, the students were asked to indicate whether their instructor was “considerate of others” or “self-centered”—these two designations representing the two extremes of the scale—and scored one and five respectively. The students were instructed to mark the position on the scale which most closely approximated their estimate of the instructor. The scales were reversed alternately in order to reduce too hurried or uncritical answers.

The instrument has only face validity and therefore can be interpreted only as the degree to which students were willing to manifest certain judgments of the instructor. However, the extreme end-points of each scale were sufficiently positive and inclusive so that, granted a minimum knowledge of ordinary common speech, students could distinguish clearly between the two extreme qualities. Administration of the scales required a minimum of questions from the students. Moreover, the students were not asked to sign their names, although they were asked to include their student numbers. They were assured that the instructor would not see any individual rating, and that the materials would be used only by the researcher, who was himself not an instructor in any of the experimental sections.

From a standpoint of pure experimental research the various errors and flaws in the design are apparent. It will not be necessary to dwell upon these but to state that, with full awareness of error, there still remain data which offer some suggestive possibilities to those interested in class-size research of the kind indicated herein. Although no attempt was made to determine the reliability of the estimates, it may be assumed that lack of reliability should increase variance within sections and hence decrease the probability of obtaining significant differences. Errors of reliability, therefore, should increase the probability of supporting the null hypothesis. If, despite the untested character of the instrument, significant differences are obtained and if these differences are in the same direction, it may be assumed that there is reason to reject the null hypothesis for this particular experiment.

With respect to reliability it should also be pointed out that this experiment was designed to obtain group, not individual, reaction. Since all groups in the experiment had an equal chance to manifest favorable or unfavorable sentiments with respect to the instructor, differences between groups, when the instructor variable is held constant, may be taken to have some relationship to the principal variable of class size. That is, the scores of the groups indicate the variation between groups with respect to their willingness to manifest certain sentiments about their instructors. It is not possible to say that these sentiments are "really" the sentiments of the individual student participants in the experiment. What the "real" sentiments of the students are, as manifested by the image of the instructor, is a problem to be left to the difficult and elusive field of attitudes and their definition. The major point to be made here is that the data obtained from this experiment offer an opportunity for testing differences between classes of students regardless of individual student variation. It should be emphasized, of course, that the data from this experiment must be read with appropriate caution and limited to the sample population and classes of students in the experiment.

The Findings

The original research design called for the administration of the instructor rating scale, as well as other instruments, three times during the term. The data to be reported herein are limited to the image of the instructor as revealed by the final instructor rating obtained from the students. The original sample population and the actual number of replies obtained on the final rating are shown in Table I.

The actual sample from which the data of this report are drawn was 613 students, or approximately 78% of the total sample of 784 students.

The method of the analysis of variance was used to test for the significance of differences in the image of the instructor reported by various sections. In order to determine if it was feasible to group the sections by type, regardless of instructor, an analysis was run for instructor differences. These differences are shown in Table II.

This table shows the number of students, the mean response to the rating scale, and the F ratio. The lower the mean response the more favorable the image of the instructor. The theoretical limits of the rating scale were 10 (most favorable) to 50 (most unfavorable). As can be seen from the mean responses, all instructors in all sections were rated quite high. That is, the responses are skewed toward the more favorable end of the scale. That the students were actually so favorable toward the instructor as the scores indicated has already been dismissed as a problem apart from this particular research. What is important in this investigation is the determination of differences between sections and instructors regardless of the piling up of scores at one end or the other of the scale. The piling up of scores, in terms of statistical treatment, insofar as it is relatively constant, as in the present case, has the practical effect of limiting the variation and reducing the size of the scale. Such an outcome serves to further reduce the probability of obtaining significant differences. Hence, the differences which were obtained should be scrutinized with some care.

It is apparent that in the large and medium lecture-discussion sections there is considerable variation among instructors. This variation, in the case of the large sections, seems to be attributable to Instructor C. In the case of the medium sections the difference is apparently attributable to Instructor B. In the other two types of sections the differences are not significant. In the case of the small "free" section there is a trend toward significance which is apparently due to a combination of Instructors B and C.

From these data it must be assumed that there are differences among instructors which are greater than chance and that it would be inadvisable to pool the sections by type. It might be argued that Instructors A and D could be treated simultaneously. However, since the size of the sample permitted, it appeared best to analyze separately the differences between the sections of each instructor. These differences are shown by Table III.

From this table it may be inferred that for this experiment there are significant differences not only among instructors but between sections, when the instructor variable is controlled. The differences, however, are not all related to the same factors. Thus, for Instructor A, the differences appear to be attributable to his more favorable image in the small lecture discussion section. For Instructor B the differences appear to be related to his less favorable image in the large lecture-discussion section. For Instructor C the differences appear to be attributed most logically to the less favorable image manifested by the medium lecture-discussion section. While for Instructor D, where the differences are not significant, although there is a trend toward significance, variation appears to be related to the more favorable image reflected ty the small lecture-discussion section.

However, certain trends appear to demand some consideration. In all cases, the most favorable image of the instructor is manifested by the small lecture discussion section. There is also a strong tendency for both the large and medium large lecture-discussion sections to produce less favorable images of the instructor than the small lecture-discussion section. A

TABLE I

SIZE OF SAMPLE BY INSTRUCTOR AND TYPE OF SECTION

	Large Lecture-Discussion Sec.		Medium Lecture-Discussion Sec.		Small Lecture-Discussion Sec.		Small "Free" Sec.	
Instructor	Enroll-ment	Present at Final Rating	Enroll-ment	Present at Final Rating	Enroll-ment	Present at Final Rating	Enroll-ment	Present at Final Rating
Instructor A	97	77	48	40	24	21	25	20
Instructor B	98	74	50	40	25	25	25	21
Instructor C	98	82	47	40	25	22	26	18
Instructor D	101	64	47	35	23	17	25	17
Totals	394	297	192	155	97	85	101	76
Percent in attendance at Final Rating	75.4%		80.7%		87.6%		75.2%	

TABLE II

IMAGE OF THE INSTRUCTOR WHEN TYPE OF SECTION IS HELD CONSTANT

		Instructor							
		A		B		C		D	
Class Type	F	No.	Mean	No.	Mean	No.	Mean	No.	Mean
Large Lecture-Discussion Section	4.3*	77	19.6	74	19.8	82	17.3	64	20.0
Medium Lecture-Discussion Section	7.5*	40	19.4	40	16.3	40	20.3	35	21.7
Small Lecture-Discussion Section	1.4	21	15.1	25	15.5	22	16.3	17	17.8
Small "Free" Section	2.0	20	19.5	21	16.7	18	16.8	17	19.5

*Differences significant beyond 1% level.

TABLE III

IMAGE OF THE INSTRUCTOR WHEN INSTRUCTOR VARIATION IS CONTROLLED

		Large Section		Medium Section		Small Section		Small "Free" Section	
Instructor	F	No.	Mean	No.	Mean	No.	Mean	No.	Mean
Instructor A	3.0*	77	19.6	40	19.4	21	15.1	20	19.5
Instructor B	9.7*	74	19.8	40	16.3	25	15.5	21	16.7
Instructor C	5.6**	82	17.3	40	20.3	22	16.3	18	16.8
Instructor D	2.1	64	20.0	35	21.7	17	17.8	17	19.5

* Differences significant beyond 5% level.
**Differences significant beyond 1% level.

combination of these trends appears to produce significant differences.

Still other possibilities are indicated by these data. Since both instructor and section differences have been demonstrated to be significant, the question arises as to whether still more important differences may be due to the interaction between instructor and student by type of section. It is possible to obtain a precise test for this hypothesis but the procedures are involved and impractical for a limited suggestive study of this type.[5] However, inspection of Tables II and III indicates the possibility that this hypothetical interaction factor may be of considerable importance. It would appear clearly that some instructors are more favorably received in some classes than in others. That is, instructors usually had a "good" or "bad" section. Hence, it is suggested from these clues that perhaps a most fruitful hypothesis for further research is that the crucial variable is the "instructor in the situation." That is, some instructors are more favorably received in certain kinds of classes than in others. This factor may also be crucial for variables other than instructor image.

One final datum was available. Each instructor was asked to give, on the basis of his best judgment, an estimate of how he felt he had been received by the students in the experiment. The results should be informative for any teacher who assumes that he can successfully take the pulse of his class. Instructor A stated the medium lecture-discussion section was best for him and that the large section was impossible. Actually, the small lecture-discussion section was his "best" and there was little difference in his reception by the others.

Instructor B stated that he was not and did not feel effective with large groups. He was right with respect to the image reflected by the large lecture-discussion section but he was actually received slightly more favorably by the medium lecture-discussion section than by the small informal "free" section.

Instructor C was, if anything, the most experienced teacher in terms of lecture and lecture-discussion techniques, although all instructors were teachers of considerable experience. This instructor indicated that he would be received most favorably in the larger sections and indicated some feelings of discomfort on his part and doubt as to his ability to handle smaller sections as successfully as larger sections. Actually he was rated most favorably by the small lecture-discussion section and most unfavorably by his medium-large lecture-discussion section.

Instructor D indicated that he preferred the small sections and assumed that he was most favorably received by them. It was true that he was most favorably received by the small lecture-discussion section but there was not a great amount of difference in his reception by the other sections, including the small informal "free" section.

Conclusions and Implications

Although it is necessary to interpret with caution the data that have been presented, it is possible to indicate certain tentative conclusions:

1. There are differences among instructors as to the way that they are received by classes of different sizes.
2. There is a tendency for instructors, other things being equal, to be more favorably received by small sections than by larger sections of the same type.
3. The informal "free" sections used in the experiment did not produce consistently the most favorable image of the instructor. When compared with the small lecture-discussion sections of approximately equivalent size it is apparent that both teaching procedures and size must be considered.
4. There is some reason to believe that the interaction factor may be crucial to the perception by the student of the instructor.
5. Finally, this experiment supports the hypothesis that instructors are not able consistently to estimate their reception by their classes, nor to distinguish correctly the difference in images reflected by different classes.

This experiment, as has been noted, is quite limited in scope and less rigorous than might be desired. The results have been presented, however, because they open certain lines of investigation which have, in general, been ignored in class size research.

Most class size studies have been confined to pupil achievement on some fixed instrument. Yet there are few teachers who would accept uncritically the hypothesis that pupil achievement on an objective scale necessarily measures all or even the most important teaching outcomes.

---

[5] The reader who is interested in this statistical test is referred to George W. Snedecor, Statistical Methods (Ames, Iowa: Iowa State College, 1946), pp. 284 ff.

This is not to say that objective measures should be discarded for subjective evaluations. This research has, itself, indicated the inability of teachers to estimate subjectively certain factors such as class response to the instructor. But it may well be argued that most measures of pupil achievement estimate only the bare skeleton of the educational process. More important behavorial changes are unmeasured and these changes constitute the essence of education.

In the process of educational change, insofar as it corresponds to personality change in general, the relationship of teacher to student has often been noted as an important factor. It has also been noted that learning is facilitated by a favorable environment. Not the least of this environment is the teacher himself. The emotional response to the teacher is an important concomitant of the learning process. Insofar as there may be a relationship between the educational process and the image of the teacher in that process, it would seem important to investigate further not only this relationship but also the effect of methods of instruction, size of class, and class-instructor interaction upon the image that students have of the instructor.

Editor's Note: Dr. Holland died in June, 1953, while on a field assignment in Havana, Cuba.

SOME EDUCATIONAL IMPLICATIONS OF THE INFLUENCE OF REJECTION ON ASPIRATION LEVELS OF MINORITY GROUP CHILDREN

REGINA M. GOFF
Morgan State College
Baltimore, Maryland

THIS STUDY is concerned with the social pressure of rejection as it acts as a barrier to effective intellectual and social functioning of minority group children. The purpose is to consider approaches in guidance which may offset attitudinal and action patterns which follow rejection and which subsequently interfere with goals or aspiration levels of the individual who is rejected.

The term is used in the sense of a feeling state which represents a response to specific overt action patterns such as being barred from activities and institutions freely attended by others. The act of rejecting is essentially a dispatching of annoyers to which the designated individual responds with discomfort. An assumption made is that attitudes have a pervasive effect on behavior; that goals personally anticipated have an “inner structure”[1] which reflects attitudes of confidence; that experiences subsequent to the original goal-setting may depress attitudes and thus deflect the individual from anticipated paths leading to the attainment of the goal. Further, ways of thinking concerning the self become habituated through time in terms of social reinforcements. The individual thus learns to place values on himself in terms of positive or negative experiencing. It is the desire here to indicate how educational experience may contribute to positive conditioning with reference to the self.

In analysis of the problem, there is presented, first, an exploration of the probable strength of intervening annoyers and interferences through an appraisal of the picture which a designated group of children have of themselves, including, among other things, their own evaluations of their abilities and the concomitant goals which they have set. This is followed by a consideration of actual achievement as it is found in adult members of the same group. The probable strength of annoyances in forestalling the realization of goals may be indicated by the consistency or discrepancy existing between goals of young group members and actual achievement as it is found in mature members of the same group. The approach thus followed includes, a presentation of original, impirical data on the children studied and a summary of the findings. Next, a comparative statement of the status of adult members of the same group, and finally interpretations and implications with reference to offsetting the effects of rejection.

Description of the Study

One hundred twenty children, including 60 boys and 60 girls were interviewed in Durham, North Carolina, in the summer of 1951. Represented in the group were 6 to 8 year old children and those aged 12 to 14. Equal members were in each group. Two age groups were used in order to note any changes in attitudes or outlook during the years of growing awareness, deepening perception, the acquisition of new knowledges, and new realizations concerning individual abilities. Upper and lower income groups are also represented in recognition of the fact that individuals in different social and economic groups “act in response to their cultural training, to their particular system of socio-economic rewards.”[2] Contrast in social experiencing and training result in subjective intellectual and emotional reactions which result in contrast in thought, feeling, and action.

Questions asked centered around successes or failures in competitive out-of-school activities (games and sports) in which they freely engaged; ranking of the self in relation to in-school academic performance; ambitions, goals or aspiration levels, and major wishes held. Where one places himself, reflects personal feelings of potentialities, confidence, and worth of self. Ambitions or goals are natural consequents of self-evaluation. That there may be wishful thinking rather than objective, realistic evalu-

---

1. Kurt Lewin, et al. "Levels of Aspiration," in Personality and Behavior Disorders, Vol. II (New York: Ronald Press, 1944), p. 335.

2. Allison Davis. "Light from Anthropology," in Cultural Groups and Human Relations, Conference on Educational Problems of Special Cultural Groups (New York: Bureau of Publications, Teachers College, Columbia University, 1951), p. 84.

ations, is a possibility. Nevertheless, such thinking furnishes a basis for action. In-school and out-of-school performances were considered to note any possible differences in ranking of the self in freely chosen activities and those which are adult imposed tasks. Wishes, revealing unfulfilled desires or "formulated aspirations" indicate sensitivity to lacks. This ego involved sensitivity provides vulnerable areas in personality structure on which negative pressures may make immediate and noticeable inroads.

The qualitative data from the interviews are categorized and later quantified in terms of percentage in various categories. In order to obtain significant differences in responses according to age level and economic bracket, where appropriate, the chi-square technique was employed.

In giving results obtained, it is recognized that 6 to 8 year old children, because of limited experience, may present fanciful conjectures. Nevertheless, if such should be the case, it is not amiss to place credence in responses inasmuch as phantasy has its origin in experience, real or vicarious, and thought patterns emerge from such content. It may be added that 12 to 14 year old children may voice social expectations or approved-of-goals rather than true individual expectancies. However, in 46% of the cases, the ambitions of children differed from parental expectations of them which might indicate some genuine personal concern on the part of the subjects.

Statement of Findings

1. Competency in in-school and out-of-school activities
   - A. In freely chosen out-of-school activities, lower income children indicated a decrease in confidence with increase in age. (Significant at the .01 level of probability.)
   - B. Feelings of inadequacy in relation to school subjects were notable in the low income group. But 33% of the girls at each age level felt competent, while 63% of the younger boys and 37% of the older had feelings of academic success.
   - C. Upper income girls indicated rising assurance with age increase in relation to both in-school and out-of-school activities. (Significant at the .02 level of probability.)
   - D. There was no significant difference in levels of confidence expressed by 6 - 8 year old, lower and upper income boys. Older upper income boys expressed more assurance in relation to school subjects than the comparable lower income group. (Significant at the .02 level of probability.) However, there was greater reluctance on the part of upper income boys to make decisive statements concerning out-of-school activities. Statements such as "Maybe I do as well," "Sometimes," were typical.
   - E. Results obtained indicate that feelings of worth exist to some extent in all groups and are found most often in younger children. In the upper income group, children rank themselves as equally competent or better than their peers more often than they rank themselves as decisively less capable.
   - F. Lower income girls hold least feelings of assurance and self-esteem.

II. Ambitions held, presented according to age, sex, and income level and in order of first three preferences
   - A. Lower income group:
     - Boys: 6 - 8—Firemen and policemen (26%); brickmasons, builders, mechanics (24%); lawyers (20%)
       12 - 14 —Big league ball players (40%); doctors (27%); contractors builders (14%)
     - Girls: 6 - 8—Teachers (34%); movie stars, dancers, models (15%), cooks (14%)
       12 - 14—Teachers (30%); movie stars, dancers, models (15%); secretarial activities (10%)
   - B. Upper income group:
     - Boys: 6 - 8—Policemen and firemen (34%); builders and mechanics (32%); agents in real-estate offices (14%)
       12 - 14—Doctors and lawyers (46%); scientists, "biologists", "chemists" (20%); architects (10%)
     - Girls: 6 - 8—Teachers (34%); nurses (26%); home makers: mothers (24%)
       12 - 14—Teachers (37%); movie stars (27%); secretaries (14%)
   - C. Seventy-nine percent of the lower income boys and 33% of the girls felt that they would reach their goals, while 53% of the upper income boys and 67% of the girls made decisive statements of assurance.
   - D. Lower income children held money and lack of opportunity as major interferences while the upper income children named bad luck, ill-health, physical disabilities, and death as possible interventions.
   - E. Lower income girls expressed least security with reference to probable success.

III. Expression of wishes presented in order of first three choices
   - A. Lower income group:
     - Boys: 6 - 8—Toys (Only wish expressed.)

  - 12 - 14—Money, clothes
- Girls: 6 - 8—Clothing, food, homes
  - 12 - 14—Improved physical appearance, homes, families, and money

B. Upper income group:

- Boys: 6 - 8—Toys, pets
  - 12 - 14—A wide spread of desires was reported from the acquisition of Eagle Scout badges, success in occupation, money for philanthropic purposes, to more sober wishes for good health to eternal life
- Girls: 6 - 8—Toys, pets
  - 12 - 14—As above, wishes were divergent with no heavy localization in any particular category. These included distribution from personal luxuries and travel to fulfillment of social needs; "places for colored children to swim," skating rinks, admittance to the "big library."

C. Particularly in the lower income group were findings similar to those of Gray[3] in which wishes were more in terms of the material and concrete than in terms of the abstract; happiness, health.

Summary of Findings

A glance at trends throughout the findings reveals, with reference to lower income girls, a straight line relationship between low confidence levels, little anticipation of success, and feelings of deprivation as revealed by wishes in areas which contribute to inner securities. Similarly, a thread of continuity appears to run through the findings with reference to the remainder of the group reporting. In general, feelings of confidence are expressed more often than not, positive attitudes of competence exist more frequently than negative ones, ambitions are directed toward occupations which yield substantial economic returns, success in areas of performance are most often expected, and wishes generally are in terms of further self-enhancement.

While these children have no doubt experienced ego-deflation in their movement in the larger world, they evidently see them as isolated from the context of life of the future. There probably exists at these ages no generalized notions of the impact which a totality of rejecting episodes of the larger world may have on the fruition of ambitions. When asked to state conditions which might interfere with the realization of goals, but two mentioned rejecting experiences and conflict in inter-group relations although a much larger number had no doubt experienced unfavorable situations.

The Adult Population

If we now turn to the adult population of Negroes in American life, we find but 2.6% of the total group in the professions with more than half that number in the teaching occupations and the remainder spread thinly as doctors, trained ministers, and lawyers. Discriminatory and rejecting practices in industry have leveled the ranks of Negro workers to unskilled and semi-skilled employment. Hardly 5% of the total Negro population can be considered in the upper income bracket. How much of the discrepancy which we note between children's confidence levels, their ambitions and wishes toward improvement and adult reflection of lack of fruition of these can be attributed to outright techniques of economic and social discrimination and rejection, or to misjudgment of ability, and how much to apathy or a giving up in the process of development is hardly discernible. Yet, the question appears worthy of consideration. It is highly probable that aspiration levels are lowered, confidences shaken, and wishes abandoned because of psychological omissions in training and guidance. It is also likely that much by way of prevention of wastage of ability and warping of personality could be achieved if, before their inception, effort were made to offset negative attitudinal patterns which are barriers to achievement. While the Negro group was used for the illustrative purposes, it would appear appropriate to assume that the same principles apply to any individuals in comparable unfavorable situations for common causes are present despite other variables.

Interpretations and Educational Implications

As indicated, lower income children showed a marked feeling of inadequacy in relation to school subjects. It is possible that attitudinal patterns stem most directly from early home and cultural influences. In other words, these children have learned a particular response pattern. They have no social heritage akin to scholarly production, nor perhaps do they even have family exposures which reflect a tradition of schooling. Such omissions generate an unfamiliarity with abstract symbols and attendant feelings of shyness concerning them. The discernible trend to feel more at ease in physical games and sports may be worthy of attention. The school might recognize the role of feelings of competence in physical activities from the point of view using these initial security reactions for the development of generalized feelings of adequacy. Praise and commendation following successful endeavor in deliberately chosen games which require thought as well as skill, with this requirement pointed out to the child, releases latent streams of inspiration which invoke motivation to tackle problems with similar requirements in other areas. A greater measure of successful anticipation will accompany effort, and early tendencies to feel incompetent in relation to diverse tasks will be eased.

---

[3] S. Gray. "Wishes of Negro Children," *Journal of Genetic Psychology*, LXIV (1944), pp. 225-237.

Incidentally, a factor influencing attitudes toward success in out-of-school activity is absence of adult domination and censure. The child, relieved of the tension of straining for goals set by adults unreservedly directs attention and enthusiasm to the experience at hand. Also, he probably chooses those activities in which he has demonstrated his competence. Teachers might well assume a “background” or “hands off” attitude, thus avoiding unnecessary intrusions which set up emotional restraints.

The discernible trend of decrease in confidence level with increase in age in lower income children has implications for the basic problem —the probable force of the impact of rejection. Greater receptivity, less resistance to social depressors characterize the individual who already has initial feelings of inadequacy and whose confidence level is continuously lowering. In such instances, there are hardly notions of success freely entertained nor high goals set with reference to the larger competitive world. Lower income girls merit special attention for their reactions indicate a greater sensitivity with reference to the self. A cue is given here for the building of thought patterns, outgrowths of experience, which negate feelings of self-limitations. Otherwise, circumstances of rejection will be accepted as a natural consequence of imagined inadequacy. A capable person may thus never discover latent abilities and as a consequent waste inherent talent.

The contrast furnished by the upper income girls, in which there is increase in self-confidence with age increase, may reflect among other things more considerate classroom treatment. Not infrequently, children from “better” families” are academically favored or given a place of importance which serves to nourish the ego and to influence effort and output. If the difference in feelings of competence can be attributed in part, to inter-personal parent-child, teacher-pupil relations, it is incumbent upon parents and teachers to begin in the early stages of the child's life, particularly low-income children, to contrive planned activities which provide opportunity for independent performances which give status in the respective groups. This in turn provides satisfactions which bolster the ego.

Older upper income boys revealing less confidence than girls in the same group, no doubt have a greater attunement with reality. There is more movement and mingling outside the home and classroom with greater opportunity for sharp person to person comparisons and increased awareness of group standards. Outlooks, as a result, are less hopeful.

As indicated, feelings of worth exist to some extent in all groups. Where found, these positive attitudes may be thought of as “security anchors” which, if strengthened, may, figuratively speaking, provide a basis for the building of a reserve of “ego energy” which in turn holds the individual fast when he is faced with annoyers designed to deflate.

It is interesting to note in relation to occupational choices, that irrespective of socio-economic level and roles of parents, ambitions of children were directed away from more menial types of employment. The increasingly large number of picture magazines and comic strips high-lighting activities of the world outside the home may have some bearing on choices made. An influence, too, may be the recent advent of a few Negroes in occupational avenues previously closed to them. An indication of social direction of a given time is revealed by the beliefs and values held by the children of the culture.

A question sometimes arises with reference to the advisability of giving the growing child a full view of the social world of which he is a part, and of his possible position and treatment within it. As stated earlier, the Negro child voiced little evidence of seeing a relationship between future success in chosen fields and his particular group membership. Is it wise to tell a child of underprivileged parents and deprived homes, or a child of a minority group that he might not be invited to play with some children whose parents consider them better; that he will not be allowed to enter some theaters and restaurants; that adult members of his group are not welcome to live in some neighborhoods and may be bombed out; that there are industries and firms which will not hire them? Is there a risk of unduly planting seeds of inferiority and antagonism? Is it better to let the individual smack up against the culture without forewarning? Perhaps either approach as stated is inadequate. However, if revealing the culture carries with it a “readiness for,” a full view of the larger world appears feasible. An approach tinged with self-pity or hatred, rather than understanding, is detrimental. However, if the growing individual were made aware of annoyances and interferences there would more likely be wholehearted cooperation in training designed to weaken the impact. Further, there would follow increased determination to succeed in desired ends.

Though recurring patterns of rejection, mobilized by the culture have persisted through time, few behavioral responses conceived of as psychological defenses have been suggested for withstanding them. However, such insightful training seems a possibility. Teacher training institutions might consider even further than at present the social and psychological role of future teachers and emphasize the meaning of social morality, the study of social-psychological problems, principles of mental hygiene and guidance, and dynamics in human relations. The examination of casual sequences in behavior, attempts to offset the effects of known but uncontrollable causes, the utilization of experimental data in the area of human relations, and use of findings from socio-psychological studies in child guidance would all appear to be aids in filling the notable gaps in the training of minority group children.

ERRATA

We regret that the following errors appeared in Mr. Harry Ruja's article "Outcomes of Lecture and Discussion Procedures in Three College Courses," published in the Journal of Experimental Education, June, 1954:

Page	Col.	Line	As Printed	Should Have Been
387	2	14	24	23
389	1	26	24	2^4
389	1	4th from bottom	to fav-	two fav-
390	1	27	$\underline{P} < .001$	$\underline{P} < .01$
393	1	4	psychology students	psychology lecture students
393	1	Ref. 1st line	Teaching Psy-	Teaching in Psy-

Table IV was omitted and should have followed page 393:

TABLE IV

SUMMARY OF COMPARISONS BETWEEN LECTURE AND DISCUSSION

	Philosophy 1A	Philosophy 1B	Psychology 1
Performance	=	=	<
Gains on Bell	=	=	=
Instructor Rating	>	>	=
Proportion of Favorable Free Comments	=	=	=
Names Known	>	>	>
Intensive Aquaintance	>	=	=

Note: The comparisons begin with the discussion section. Thus, the first cell in the upper left corner is to be read thus: The discussion section of Philosophy 1A equalled the lecture section in performance.

Journal of Experimental Education

Volume XXIII March, 1955 Number 3

STATISTICAL ANALYSIS OF SCORES FROM COUNTERBALANCED TESTS

JULIAN C. STANLEY[1] *
University of Wisconsin

478 V

The Simplest "Crossover" or "Switchover" Design

MANY psychological studies deal with the comparability of mean scores on various tests, such as IQ's for the Revised Stanford-Binet and the Wechsler Intelligence Scale for Children (WISC). If a random half of the available subjects are given the S-B and the other half the WISC under carefully equated conditions, a t-test for the significance of the difference between the uncorrelated means provides an answer to the question posed.

However, the use of independent groups may be quite inefficient, since the correlation between test scores for the same individuals will often be high. Testing only one group of children with both instruments and then using analysis of variance significance tests for correlated observations should be more efficient. This method causes some difficulties in design and analysis not encountered with the independent-group technique, but these are not serious enough to disqualify it.

The simplest design of this sort occurs when only two tests are being compared. These two tests should be administered in counter-balanced order to random halves of the total group. Half of the subjects should be tested first with, say, the S-B and second with the WISC; the other half should have the WISC followed by the S-B. In this way sequence effects will not be confounded with the difference in difficulty of the two tests. Taking a subject at random from the S-B, WISC group and one at random from the WISC, S-B group results in the simple latin square of Table I. There will be half as many such latin squares as there are subjects in the investigation, or as many as there are individuals in either group. It is more convenient to set up the scores for analysis in the form of Table II, where the two orders are kept separate. This is a consolidation of n latin squares like the one shown in Table I.

The Table II scores can be analyzed for orders, individuals, tests, and sessions (first test versus second test) by a method set forth in considerable detail by Lindquist (17:273-81), Edwards (3:319-27; 4), Walker and Lev (24: 373-81), and Grant (9:428-32;10). Apparently the design has been applied to this type of problem only by Gellerman (7), who extended it by matching his two order groups individual by individual (1 with $n + 1$, 2 with $n + 2$, . . . , n with $2n$) on other variables but failed to take this additional source of correlation into account when analyzing the scores.

There are the following degrees of freedom for sums of squares from Table II: 1 each for tests, orders, and sessions; $2(n - 1)$ for differences between individuals of the same order; and $2[(n - 1)(2 - 1)]$ for the "residual within individuals," which is the individuals of the same order by tests interaction pooled for the two orders. These *d.f.* add to $4n - 1$, or $N - 1$, the usual sum. The proper "error" term for orders is the "differences between individuals of the same order" mean square. For sessions and tests the "residual within individuals" mean square is the appropriate divisor. The reader interested in the rationale of the analysis should consult Cochran (2), Edwards (3, 4), Grant (9, 10), Kempthorne (13:Ch. 29), and Lindquist (17).

Some examples of counterbalanced designs involving two tests are the studies of Frandsen and Higginson (6), Gerboth (9), Holland (12), Krugman and others (15), Kutash (16), Pastovic and Guthrie (19), Peel (20), and Vanderhost and others (23). Hays and Schneider (11) employed a more complex counterbalancing procedure with 16 independent groups. With the single exception of Gellerman, noted above, it appears that no *test* investigator has used the general type of design shown in Table II, though in at least eight different studies it would probably have been more appropriate than the statistical analysis reported. Since no textbook or article makes thoroughly explicit the procedure for this particular psychometric design, let us work through some data based upon the Frandsen-Higginson (6) report, showing both the correct and incorrect way to analyze it. The writer is indebted to Professor Arden N. Frandsen for these scores.

---

*Footnotes will be found at the end of the article as well as the bibliography.

The Frandsen-Higginson Study

Originally, Higginson tested 54 fourth-grade children, 27 with the S-B first and the WISC second and the other 27 with the WISC first and the S-B second. Order and score information

TABLE I

THE ELEMENTAL LATIN SQUARE IN A COUNTERBALANCED TESTING STUDY

		Tests	
Orders	Individuals	S-B	WISC
S-B, WISC	A	S-B First	WISC Second
WISC, S-B	B	S-B Second	WISC First

TABLE II

COUNTERBALANCED TESTING WITH THE S-B AND WISC

Orders		IQ's	
(Groups)	Individuals	S-B	WISC
S-B, WISC	1	First test	Second test
	2	First test	Second test
	.	.	.
	n	First test	Second test
WISC, S-B	n + 1	Second test	First test
	n + 2	Second test	First test
	.	.	.
	2n	Second test	First test

for 51 of the 54 testees was furnished the writer; for three persons the order of testing was lost at the time of inquiry. By discarding another testee randomly the writer had 50 individuals left, making a total of 100 scores, as outlined in Table III. Table IV shows how analyzing the orders separately is related to the appropriate overall analysis, which permits tests of both orders and sequences, as well as of individuals (nearly always significant) and tests.[2]

The chief finding is an $\underline{F}$ of 6.77 for tests, which with 1 and 48 degrees of freedom is significant between the 5 and 1 percent levels, IQ's being higher on the S-B than on the WISC. If the 25 scores for the S-B taken first are tested against the 25 for the WISC taken first (a simple randomized design), the $\underline{F}_{1,48} = 641/177 = 3.6$, a considerable loss of precision.[3]

The mean square of 25.00 for sequences in Table IV is less than the residual error term, which is hardly surprising because the average practice effect inferred from Table III means is -1.00 IQ points. Also, the two order (group) means, 106.38 and 102.74, do not differ significantly when considered in terms of variation among individuals within orders.

The six sums of squares are secured as follows from the figures in Table III:

$$\underline{SS_o} = \frac{1}{100}\left\{2\left[(5319)^2+(5137)^2\right]-(10{,}456)^2\right\} = 331.24.$$

$$\underline{SS_i} = \frac{1}{50}\left\{25(2{,}212{,}516)-\left[(5319)^2+(5137)^2\right]\right\} = 12{,}647.40.$$

$$\underline{SS_{tests}} = \frac{1}{100}\left\{2\left[(5316)^2+(5140)^2\right]-(10{,}456)^2\right\} = 309.76.$$

$$\underline{SS_s} = \frac{1}{100}\left\{2\left[(5253)^2+(5203)^2\right]-(10{,}456)^2\right\} = 25.00.$$

$$\underline{SS_r} = \frac{1}{50}\left\{50(574{,}768+534{,}020)-25(2{,}212{,}516)-2\left[(2716)^2+(2603)^2+(2600)^2+(2537)^2\right]+\left[(5319)^2+(5137)^2\right]\right\} = 2195.24.$$

$$\underline{SS_{total}} = \frac{1}{100}\left[100(574{,}768+534{,}020)-(10{,}456)^2\right] = 15{,}508.64.$$

As Grant (9:430) has pointed out, when only two tests are being compared, each of the three factors (tests, orders, and sequences) is completely confounded with the interaction of the other two. Therefore, the sum of squares for "tests" is identical with that for the orders $\underline{x}$ sequences interaction:

$$\frac{1}{100}(2716+2600-2603-2537)^2 = 309.76.$$

Also, the sum of squares for "orders" is the same as the sum of squares for the tests $\underline{x}$ sequences interaction:

$$\frac{1}{100}(2716+2603-2600-2537)^2 = 331.24.$$

Likewise, the "sequences" sum of squares may be secured via the orders $\underline{x}$ tests interaction:

$$\frac{1}{100}(2716+2537-2603-2600)^2 = 25.00.$$

A little reflection will convince the reader that these relationships are quite reasonable. For example, the orders $\underline{x}$ sequences interaction will be significant only if the mean of the S-B taken first minus the mean of the WISC taken second differs significantly from the mean of the WISC taken first minus the mean of the S-B taken second:

$(\underline{M}_{S\text{-}B_1} - \underline{M}_{WISC_2}) - (\underline{M}_{WISC_1} - \underline{M}_{S\text{-}B_2})$. This may also be written

$(\underline{M}_{S\text{-}B_1} + \underline{M}_{S\text{-}B_2}) - (\underline{M}_{WISC_1} + \underline{M}_{WISC_2})$, showing that the $\underline{o}$ $\underline{x}$ $\underline{s}$ interaction is equivalent to the "tests" main effect (S-B versus WISC) when only two tests are compared.

If just the "tests" effect is of interest, as in the Frandsen-Higginson study, the mean difference between S-B and WISC scores of the 50 testees may be tested for significant deviation from zero somewhat more easily than in Table IV. Obtain the 25 differences (say, S-B - WISC) for each order and the sum and sum of squares for each of the two sets of differences. Then

$$\underline{F}_{1,48} = \frac{\dfrac{\left(\sum\limits_{1}^{50}\underline{D}\right)^2}{50}}{\dfrac{\frac{1}{25}\left[25\sum\limits_{1}^{50}\underline{D}^2-\left(\sum\limits_{1}^{25}\underline{D}\right)^2-\left(\sum\limits_{26}^{50}\underline{D}\right)^2\right]}{48}}$$

$$= \frac{619.52}{91.47} = 6.77.$$

This is identical with the $\underline{F}$ for "tests" in Table IV, but note that both numerator and denominator of the $\underline{F}$-ratio are twice as large as the analogous mean squares of Table IV.

The statistical procedure used most often in studies employing two counterbalanced tests is a $\underline{t}$-test for the significance of the difference be-

TABLE IV

COMPOSITE ANALYSIS OF VARIANCE OF FRANDSEN-HIGGINSON (6) DATA

Source of Variation	Sum of Squares			Degrees of Freedom			Mean Square		
	S-B First	WISC First	Overall	S-B First	WISC First	Overall	S-B First	WISC First	Overall
Orders ($t \times s$)	....	....	331.24	....	....	1	....	....	331.24 $F = 1.26$ ƒ
Individuals	7,472.28 +	5,175.12 =	12,647.40	24 +	24 =	48	311.34*	215.63*	263.49* $F = 5.76$
Tests ($o \times s$)	255.38 +	79.38 ≠	309.76	1 +	1 ≠	1	255.38	79.38	309.76** $F = 6.77$
Sequences ($t \times o$)	....	....	25.00	....	....	1	....	....	25.00
Residual ($i \times t$)	1,492.12 +	703.12 =	2,195.24	24 +	24 =	48	62.17	29.30	45.73
Total	9,219.78 +	5,957.62 ≠	15,508.64	49 +	49 ≠	99	....	....	....

* $P < .001$

** $.01 < P < .05$

ƒ $\frac{331.24}{263.49}$

tween the correlated means, ignoring both order and sequence. This amounts to lumping the order sum of squares with the sum of squares for individuals and putting the sum of squares for sequences into the residual term. In the Frandsen-Higginson investigation such consolidation makes little difference, since neither orders nor sequences contribute much variation. Where the two groups are randomly selected halves of the total sample utilized in the study, the order effect will probably be great only if experimental controls differ considerably for the two groups or the two tests are quite different in content or difficulty, so that taking a particular one first facilitates work on the second test much more than taking the other test first would do. When this latter condition holds, the analysis of variance is vitiated (see 13, Ch. 29). As will be seen later in this paper (Table V), sequence (largely practice) effect may be a huge source of variation.

Estimation of Test Means

The counterbalanced design of the Frandsen-Higginson (6) study permits us to estimate that for this fourth-grade group the S-B IQ's run $106.32 - 102.80 = 4.52$ points higher than WISC full-scale IQ's. Since practice effect is negligible, these two means may be taken to be equivalent to the means that would have been secured had all 50 subjects been tested with one *or* the other of the tests. When practice effect is significant, however, the test means should not be accepted as such estimates without a suitable correction, although the *difference* between them may still be used directly. For instance, Krugman, Justman, Wrightstone, and Krugman (5: 477) in a counterbalanced study commented that "At almost every age level, the obtained mean IQ's are above 100, indicating that the sample chosen was slightly above average, although it was hoped that an average sample would be obtained by selecting representative schools," but they did not mention practice effect as a possible biasing factor.

Estimation of r's

Coefficients of correlation between the two tests should be based upon within-order sums of squares and sums of products if either order differences or sequence effects are appreciable. Group differences tend to increase the r when computed from the overall scores, while practice effect lowers it. In the Frandsen-Higginson (6) study the various r's are: .7069 for the "S-B first" group; .7678 for the "WISC first" group; .7329 overall; and .7285 within groups.[4] Thus in this instance the two methods make no difference whatsoever in the second decimal place, but there are probably many situations where the overall r will be considerably different from the within-groups r. Illustrations of this discrepancy will be given in Table VI. If X represents any score on one test and Y any score on the other, the Frandsen-Higginson within-groups r may be computed directly as shown at the top of page 9.

Estimation of Standard Deviations

Standard deviations will be affected by differences between groups and by practice effect, being increased by both. An unbiased estimate of the population variance of the S-B IQ's may be secured by taking the 232,544 figure in the denominator of the computation at the top of the next page and dividing it by 2(24)(25), or 1200. The divisor in general is $2(n - 1)(n)$, where n represents the number of individuals in either group. To obtain an approximately unbiased estimate of the population standard deviation, extract the square root of the variance.

To recapitulate: Counterbalanced testing yields a direct estimate of differences between forms or tests, since each is affected about equally by practice effect, but the counterbalanced design calls for revised methods for computing coefficients of correlation and standard deviations if practice effect is appreciable and for a more elaborate significance test.

More Than Two Tests or Forms

Three Forms of the Same Test and Three Orders

During the process of equating an old form (*A*) of a certain aptitude test with two new forms (*B* and *C*), the writer and several assistants tested 30 eleventh and twelfth grade boys in a state orphans' home in three counterbalanced orders: ABC, CAB, and BCA. Orders CBA, BAC, and ACB were not used. Ten boys, six from the eleventh grade and four from the twelfth, were assigned at random to each order. One test was given each day for three consecutive days under carefully standardized conditions. Table V shows the analysis of variance of the 90 scores. Individual, form, and day differences are all significant well beyond the .001 level, while the F ratio of 199.54/345.75 for order means is not significant.

Some investigators would analyze this data by ignoring the "order" classification entirely and performing a two-way analysis of variance (individuals and forms). Order itself is probably of little importance in most well controlled studies. However, when practice effect (days) is a large source of variation, as it is here, the significance test for the differences among form

TABLE V

ANALYSIS OF VARIANCE OF THREE FORMS OF A TEST ADMINISTERED IN THREE COUNTERBALANCED ORDERS TO 30 DIFFERENT INDIVIDUALS, ONE FORM EACH DAY FOR THREE DAYS

Source of Variation	Sum of Squares	*d.f.*	Mean Square
Orders	399.0889	2	199.54
Individuals within orders	9335.2333	27	345.75*
Forms	4361.6889	2	2180.84*
Days (Practice effect)	3509.3556	2	1754.68*
Forms × days interaction within orders**	6.1556	2	3.08
Residual (individuals within orders × forms interaction)ƒ	2373.4667	54	43.95
Total	19984.9889	89	

* $P < .001$

**For discussions of this term see Lindquist (17:278-281) and Edwards (3:324-326; 4). Lindquist analyzes the forms × days interaction sum of squares into two independent components, forms × days interaction between orders, which is equivalent to the "orders" effect above, and forms × days interaction within orders. Edwards divides the orders × forms interaction into days and a "Latin square error" remainder (4:126); he concludes that latin square error is not likely to be a significant source of variation (3:326; fn. 10). Grant (9:435-437) calls it "square uniqueness," saying that it may be tested to determine whether or not "it has been inflated or deflated by the unique pattern of confounding which occurred in the interactions of this particular square." (P. 435).

ƒ The residual sums of squares for the three orders, each with 18 *d.f.*, are: ABC, 618.53; CAB, 507.53; and BCA, 1247.40. Bartlett's test of homogeneity (21:250) yields a chi-square value of 4.22 with 2 *d.f.*, to which corresponds a *P* of approximately .15.

TABLE VI

DISCREPANCIES BETWEEN r's AND BETWEEN VARIANCES COMPUTED FROM OVERALL DATA (ORDERS IGNORED) AND FROM POOLED WITHIN-ORDERS FIGURES FOR 30 TESTEES TO WHOM WERE ADMINISTERED TEST FORMS A, B, AND C

Statistics		Overall	Within Orders
r_{AB}		.42	.70
r_{AC}		.49	.81
r_{BC}		.40	.62
Variance of Form:	A	186	179
	B	200	154
	C	153	101

TABLE VII

A GRECO-LATIN SQUARE DESIGN, WITH n INDIVIDUALS PER GROUP (TEST FORMS A, B, AND C PRINTED ON GREEN, WHITE OR YELLOW PAPER)

Group	First Session	Second Session	Third Session
1	AW	CG	BY
2	BG	AY	CW
3	CY	BW	AG

$$r_{within} = \frac{25\sum_{1}^{50} XY - \left(\sum_{1}^{25} X \sum_{1}^{25} Y + \sum_{26}^{50} X \sum_{26}^{50} Y\right)}{\sqrt{\left\{25\sum_{1}^{50} X^2 - \left[\left(\sum_{1}^{25} X\right)^2 + \left(\sum_{26}^{50} X\right)^2\right]\right\}\left\{25\sum_{1}^{50} Y^2 - \left[\left(\sum_{1}^{25} Y\right)^2 + \left(\sum_{26}^{50} Y\right)^2\right]\right\}}}$$

$$= \frac{25(551,864) - \left[(2716)(2603) + (2600)(2537)\right]}{\sqrt{\left\{25(574,768) - \left[(2716)^2 + (2600)^2\right]\right\}\left\{25(534,020) - \left[(2603)^2 + (2537)^2\right]\right\}}}$$

$$= \frac{130,752}{\sqrt{(232,544)(138,522)}} = .73.$$

means would lose much power, since the sums of squares for practice effect and "forms *x* days interaction within orders" would be pooled with the residual sum of squares to form a new error term which is too large. In Table V the *F* for forms is 2180.84/43.95 = 49.6, with 2 and 54 *d.f.*, while the inappropriate ratio is

$$\frac{2180.84}{(3509.3556 + 6.1556 + 2373.4667)/(2 + 2 + 54)} =$$

$$\frac{2180.84}{101.53} = 21.5,$$

with 2 and 58 *d.f.* Of course, in this particular study either *F* (49.6 or 21.5) is significant far beyond the .001 level, but in some situations the difference would be important.

The reader will recall that in the Frandsen-Higginson study coefficients of correlation and variances differed little, whether computed from within orders or overall, since the order means were not far apart and practice effect was negligible. A glance at Table VI will show that this is not always true, however, since the *r*'s and variances for the 30 testees show considerable systematic discrepancies, the within-orders *r*'s being uniformly higher than the overall figures and the within-orders variances lower.

All Possible Orders

Grant (9:432-434) discusses an experiment in which all six permutations (orders) of three things were employed, stating that "when complete sets of permutations are used, all interactions within such sets of squares cancel out so that they are not confounded with the main effects" (p. 434). If Form *A* in the study summarized in Table V had not had a shorter time limit than Forms *B* and *C*, it would have been feasible to use five subjects for each of the six possible orders and to test all 30 individuals in the same room with the three forms during a single day. Each person could be assigned randomly to a seat in order to prevent inter-pair correlation caused by friendship pairing. Cheating would be minimized so much that subjects could sit in adjacent seats, rather than only in alternate ones. Experimental conditions would be controlled better than when each order group was tested in a separate room. It would even be feasible to use the 24 permutations of four forms or tests if enough subjects (at least 48) and a large enough room were available; two subjects per order yield 1 × 3 × 24 = 72 *d.f.* for testing the "forms" and "sequences" effects.

Greco-Latin Crossover Designs

Walker and Lev (24:377-381) illustrate the addition of another superimposed variable to form a greco-latin square. This is not possible with a 2 × 2 design; at least a 3 × 3 setup is required. Suppose that in the three-form equating study three different hues of paper—green (G), white (W), and yellow (Y)—had been used to make the design of Table VII, where *A*, *B*, and *C* represent the three different forms of the test. Each letter occurs once in each row, once in each column, and once with each other letter, assignment being random within the limits of these restrictions. Preferably, assignment of individuals to groups would also be random in the types of studies discussed here. The degrees of freedom are: sessions, groups,

TABLE VIII

A FACTORIAL CROSSOVER GRECO-LATIN DESIGN

School Grade or Class	Sex	Orders	Individual	I	II	III
U	M	A1 B2 C3	1			
			2			
		.	.			
		B1 A3 C2	71			
			72			
	F	A1 B2 C3	73			
			74			
		.	.			
		B1 A3 C2	143			
			144			
V	M	A1 B2 C3	145			
			146			
		.	.			
		B1 A3 C2	215			
			216			
	F	A1 B2 C3	217			
			218			
		.	.			
		B1 A3 C2	287			
			288			
W	M	A1 B2 C3	289			
			290			
		.	.			
		B1 A3 C2	359			
			360			
	F	A1 B2 C3	361			
			362			
		.	.			
		B1 A3 C2	431			
			432			

TABLE IX

DEGREES OF FREEDOM FOR THE VARIOUS SUMS OF SQUARES FROM TABLE VIII

Total between individuals:	$432 - 1 = 431$
Orders:	$36 - 1 = 35$
Sexes:	$2 - 1 = 1$
Grades:	$3 - 1 = 2$
Grade × sex × order:	$(3 - 1)(2 - 1)(36 - 1) = 70$
Grade × sex:	$(3 - 1)(2 - 1) = 2$
Grade × order:	$(3 - 1)(36 - 1) = 70$
Sex × order:	$(2 - 1)(36 - 1) = 35$
Individuals within orders:	$3(2)(36)[(2 - 1)] = 216$
Total within individuals:	$(432 - 1)(3 - 1) + (3 - 1) = 2(432) = 864$
Roman numerals:	$3 - 1 = 2$
Latin letters:	$3 - 1 = 2$
Arabic numerals:	$3 - 1 = 2$
Grade × sex × roman:	$(3 - 1)(2 - 1)(3 - 1) = 4$
Grade × sex × latin:	$(3 - 1)(2 - 1)(3 - 1) = 4$
Grade × sex × arabic:	$(3 - 1)(2 - 1)(3 - 1) = 4$
Grade × roman:	$(3 - 1)(3 - 1) = 4$
Sex × roman:	$(2 - 1)(3 - 1) = 2$
Grade × latin:	$(3 - 1)(3 - 1) = 4$
Sex × latin:	$(2 - 1)(3 - 1) = 2$
Grade × arabic:	$(3 - 1)(3 - 1) = 4$
Sex × arabic:	$(2 - 1)(3 - 1) = 2$
Individuals × roman numerals within orders in grade-sex groups:	$3(2)(36)[(2 - 1)(3 - 1)] = 432$
Residual order × roman interaction within grade-sex groups:	$3(2)[(36 - 1)(3 - 1) - (3 - 1) - (3 - 1)] = 396$
Total:	$3(2)(36)(2)(3) - 1 = 1295$

forms, and hues, 2 each; individuals within groups, $3(n - 1)$; residual (individuals x sessions interaction within groups), $3[(n - 1)(3 - 1)]$; and total, $9n - 1$. The interested reader is referred to Walker and Lev (24:378-381) for computational details, which differ from the latin-square design only in that the extra effect, hue, uses up the two d.f. which in a replicated 3 × 3 latin square would be assigned to "latin square error." Before anyone attempts to design studies of this type he should be familiar with latin and greco-latin square designs as explained by Grant (9), Edwards (3:303-332; 4), Lindquist (17:258-265), Walker and Lev (24:373-381), and Archer (1).

There are 36 different rows possible for the 3 × 3 greco-latin square design. (In the types of studies discussed here these yield only 12 different 3 × 3 greco-latin squares, since the order of the rows within a particular square is immaterial.) A completely counterbalanced design can be obtained with just 36 subjects assigned randomly to the 36 rows. This gives the following d.f., using the terminology of Table VII: days, forms, and hues, 2 each; between individuals, 35; residual, $(35)(2) - 2 - 2 = 66$; total, 107. The basic 36-person design may be replicated for the two sexes and in, say, two different school grades, yielding a total of $(36)(3)(2)(2) - 1 = 431$ d.f.: 2 each for days, forms, and hues; 140 for "between individuals within groups"; 1 each for sex and grade; 1 for the sexes x grades interaction; 2 each for sexes x days, sexes x forms, sexes x hues, grades x days, grades x forms, grades x hues, sex x grades x days, sex x grades x forms, and sex x grades x hues; and 264 for the residual.[5] The confounding influence of ten other interactions has been counterbalanced out by using the 36-orders design.[6]

A test publisher might employ this kind of design with three different types of answer sheets for three different test forms at several different age levels, in each of which the testees are subdivided according to socio-economic class.

Probably in most comparisons of two individual tests, "tester" should be introduced as an explicit variable, analogous to "sex" or "schools." If three tests are utilized, however, three testers may be made a "greek" variable, like hue in Table VII. Since tester and test are likely to interact considerably, it seems desirable that all 36 possible greco-latin orders be used.

Replication Within Orders

Obviously, it is possible to use more than one individual for each order within each grade-sex group and thereby get within-order variation among individuals, as outlined in Tables VIII and IX. If just two individuals are tested with each of the 36 possible greco-latin orders for each sex and within each of three grades, there will be $2 \times 36 \times 2 \times 3 = 432$ subjects needed.[7] The between-individuals-within-orders mean square has 216 d.f. The individuals x roman numerals interaction within orders has $3(2)(36)[(2 - 1)(3 - 1)] = 432$ d.f.; the order by column residual within each grade-sex group has $(36 - 1)(3 - 1) - 2 - 2 = 66$, which when summed over both sexes and all three grades yields 396. Presumably this "greco-latin square error" or "square uniqueness" will not be significant when tested against the individuals x roman numerals interaction, since all 12 possible greco-latin squares are involved in each of the six grade-sex groups, so its sum of squares can be combined with those for individuals x roman numerals interaction to yield a "within" error estimate with $396 + 432 = 828$ d.f.

This design illustrates replication both by using the same square and by employing different ones, thus combining the two types of designs set forth by Edwards (3, 4) and going beyond the ones that Archer (1) proposes.

The Krugman Study

In a counterbalanced comparison of the Revised Stanford-Binet with the WISC, Krugman, Justman, Wrightstone, and Krugman (15) used various examiners to test both boys and girls at ten different age levels, but their analysis took into account only age (and not in the manner discussed here). A rigorous overall design for this type of study seems feasible. It would be complex, however, for there are 24 testable main effects and interactions in addition to those involving the testing order (S-B first vs. WISC first). Furthermore, if testers were selected randomly from a defined population rather than only on the basis of availability, serious mixed-model problems might occur.

There is a limit to the extent that crossover designs may be elaborated without causing grave difficulties in manipulation and interpretation, but the writer's survey of the relevant journal literature convinces him that this limit has not been approached closely yet by most psychometricians. In particular, the 2 × 2 latin and the 3 × 3 completely permuted greek crossover designs seem to merit considerable use.[8]

Summary

Psychometricians seldom use crossover (switchover) analysis of variance procedures, even when they counterbalance tests. The simplest of these designs, a 2 × 2 comparison of the Stanford-Binet and the WISC, is illustrated in detail with scores from the Frandsen-Higginson (6) study. Then a 3 × 3 design, involving hitherto unpublished scores from three forms of an aptitude test, is analyzed. Completely permuted latin and greco-latin designs, with factorial extensions, are discussed. The writer concludes that crossover designs should be quite useful for comparing different tests or several forms of a test. He particularly emphasizes the 3 × 36 greco-latin arrangement.

BIBLIOGRAPHY

1. Archer, E. J. "Some Greco-Latin Analysis of Variance Designs for Learning Studies," *Psychological Bulletin*, XLIX (1952), pp. 521-537.
2. Cochran, W. G., and Cox, Gertrude M. *Experimental Designs* (New York: John Wiley and Sons, Inc., 1950).
3. Edwards, A. L. *Experimental Design in Psychological Research* (New York: Rinehart and Co., 1950).
4. Edwards, A. L. "Homogeneity of Variance and the Latin Square Design," *Psychological Bulletin*, XLVII (1950), pp. 118-129.
5. Fisher, R. A. *Statistical Methods for Research Workers*, 10th Ed. (London: Oliver and Boyd, 1948).
6. Frandsen, A. N., and Higginson, J. B. "The Stanford-Binet and the Wechsler Intelligence Scale for Children," *Journal of Consulting Psychology*, XV (1951), pp. 236-238.
7. Gellerman, S. W. "Forms I and II of the Arthur Performance Scales for Mental Defectives," *Journal of Consulting Psychology*, XVI (1952), pp. 127-131. Also see Stanley, J. C., "Gellerman's Complex Crossover Design," *Journal of Consulting Psychology*, XVIII (October 1954).
8. Gerboth, Renate. "A Study of the Two Forms of the Wechsler-Bellevue Intelligence Scale," *Journal of Consulting Psychology*, XIV (1950), pp. 365-370.
9. Grant, D. A. "The Latin Square Principle in the Design and Analysis of Psychological Experiments," *Psychological Bulletin*, XLV (1948), pp. 427-442.
10. Grant, D. A. "The Statistical Analysis of a Frequent Experimental Design," *American Journal of Psychology*, LXII (1949), pp. 119-122.
11. Hays, W., and Schneider, B. "A Test-Retest Evaluation of the Wechsler Forms I and II with Mental Defectives," *Journal of Clinical Psychology*, VII (1951), pp. 140-143.
12. Holland, G. A. "A Comparison of the WISC and Stanford-Binet IQ's of Normal Children," *Journal of Consulting Psychology*, XVII (1953), pp. 147-152.
13. Kempthorne, O. *The Design and Analysis of Experiments* (New York: John Wiley and Sons, Inc., 1952).
14. Kogan, L. S. "Variance Designs in Psychological Research," *Psychological Bulletin*, L (1953), pp. 1-40.
15. Krugman, Judith I., and others. "Pupil Functioning on the Stanford-Binet and the Wechsler Intelligence Scale for Children," *Journal of Consulting Psychology*, XV (1951), pp. 475-483.
16. Kutash, S. B. "A Comparison of the Wechsler-Bellevue and Revised Stanford-Binet Scale for Adult Defective Delinquents," *Psychiatric Quarterly*, XIX (1945), pp. 677-685.
17. Lindquist, E. F. *Design and Analysis of Experiments in Psychology and Education* (Boston: Houghton Mifflin Co., 1953).
18. McNemar, Q. "On the Use of Latin Squares in Psychology," *Psychological Bulletin*, XLIX (1951), pp. 398-401.
19. Pastovic, J. J., and Guthrie, G. M. "Some Evidence on the Validity of the WISC," *Journal of Consulting Psychology*, XV (1951), pp. 385-386.
20. Peel, E. A. "A Note on Practice Effects in Intelligence Tests," *British Journal of Educational Psychology*, XXI (1951), pp. 122-125.
21. Snedecor, G. W. *Statistical Methods*, 4th Ed. (Ames, Iowa: Iowa State College Press, 1946).
22. Tukey, J. W. "One Degree of Freedom for Non-Additivity," *Biometrics*, V (1949), pp. 232-242.
23. Vanderhost, Leonette, and others. "Performance of Mental Defectives on the Wechsler-Bellevue and the WISC," *American Journal of Mental Deficiency*, LVII (1953), pp. 481-483.
24. Walker, Helen M., and Lev, J. *Statistical Inference* (New York: Henry Holt and Co., 1953).

FOOTNOTES

[1] The writer is indebted to Mr. V. N. Amble and Professor Oscar Kempthorne for making many helpful suggestions concerning an earlier draft of this paper. They are not, of course, responsible for any errors or ambiguities that remain.

[2] Edwards (4) points out that the two residual terms should be tested for homogeneity before their sums of squares are combined.

62.17/29.30 = 2.12, to which corresponds a P of approximately 2 × .04 = .08. Similarly, the two sums of squares for individuals may be pooled, since $F_{\underline{24,24}}$ = 7,472.28/5,175.12 = 1.44, while the F needed for significance at the 10 percent level is 1.98. The cautious investigator may want to use Tukey's (22) test for non-additivity of row and column effects, also.

[3] The standard deviation of the "S-B first" scores (15.45, shown in Table III) is nearly significantly larger than the "WISC first" standard deviation (10.08), P being .06.

[4] Of course, before a pooled within-groups r is computed, the individual within-group r's should be tested for homogeneity. Here the difference between .7069 and .7678 yields a CR of only 0.5 via Fisher's z-transformation (5:203-204). If individuals are assigned to the various order-groups randonly, significant differences among comparable within-group r's seem unlikely, though their possibility should be recognized.

[5] This complete design is presented and discussed in the Appendix at the end of this article. Also see Archer (1:532-536).

[6] When all possible permutations are built into latinized, greco-latinized, or hyper-greco-latinized designs, the objection to the use of latin squares in psychology that McNemar (18) raised is obviated. (For a comment on his article see Kogan [14:26-27].) In testing studies, completely counterbalanced designs are often fairly easy to manage and more economical than other procedures. However, Mr. V. N. Amble has pointed out to the writer that certain procedures other than full permutation may be at least as effective.

[7] Two school grades or classes require 288 individuals, each of whom takes three different tests or forms. These may seem to be large figures to the independent investigator, but they are small compared with the numbers involved in standardization studies by major test companies.

[8] In a personal communication to the writer, Professor Edward E. Cureton points out the need for a clear discussion concerning the applicability of analysis of variance methods to those cases where one criterion of classification involves classes of the variable rather than classes of the examinees, especially when these classes are tests such as the S-B and the WISC or "comparable" forms of a test. In part, this is the repeated-measurements problem dealt with by Leonard S. Kogan ("Analysis of Variance—Repeated Measurements," Psychological Bulletin, XLV (1948), pp. 131-143) and Jack Block and others ("Testing for the Existence of Psychometric Patterns," Journal of Abnormal and Social Psychology, XLVI, 1951, pp. 356-359). Frederic M. Lord (The Standard Errors of Various Test Statistics When the Test Items are Sampled, Educational Testing Service, Princeton, New Jersey, December 1953) has examined in detail the case where a large number of forms of the same test are administered to the same group of examinees, each form consisting of a sample of items drawn randomly from a common pool of items, and has derived six large-sample standard-error formulas for this situation. However, in the illustrations used in the present study the different tests are not randomly parallel. Furthermore, random sampling of items might yield forms that when used with various groups of examinees consistently give rather different scores, especially if the number of items in each form is small. These forms could not be used interchangeably without a transformation, even though they might be "comparable" in the randomly parallel sense.

"The 'Equating' of Non-Parallel Tests" is discussed by William H. Angoff (SR-54-13, Educational Testing Service, Princeton, New Jersey, May 1954).

APPENDIX

(To Supplement Table III of the Article)
100 IQ SCORES FOR THE 50 FRANDSEN-HIGGINSON SUBJECTS

S-B First, WISC Second				WISC First, S-B Second			
Pupil No.	S-B IQ	WISC FS IQ	Pupil Sums	Pupil No.	S-B IQ	WISC FS IQ	Pupil Sums
1	132	120	252	26	118	125	243
2	121	120	241	27	118	115	233
3	112	117	229	28	117	114	231
4	138	117	255	29	113	111	224
5	125	116	241	30	123	109	232
6	131	116	247	31	114	109	223
7	117	112	229	32	106	107	213
8	96	111	207	33	101	107	208
9	121	110	231	34	101	107	208
10	112	110	222	35	112	107	219
11	85	108	193	36	100	104	204
12	120	105	225	37	103	103	206
13	115	105	220	38	96	101	197
14	124	104	228	39	96	101	197
15	112	101	213	40	120	100	220
16	91	101	192	41	104	99	203
17	96	99	195	42	103	99	202
18	91	98	189	43	107	98	205
19	91	98	189	44	112	96	208
20	97	97	194	45	103	93	196
21	110	96	206	46	82	91	173
22	106	92	198	47	95	88	183
23	97	85	182	48	85	87	172
24	92	83	175	49	80	84	164
25	84	82	166	50	91	82	173

(To Supplement Table V of the Article)

90 SCORES FOR THE 30 INDIVIDUALS TESTED WITH THREE FORMS OF AN APTITUDE TEST

Testing Order	Individual Number	Form			Sums
		A	B	C	(A + B + C)
ABC	1	70	54	58	182
	2	69	70	62	201
	3	69	44	60	173
	4	67	60	65	192
	5	56	57	52	165
	6	52	58	57	167
	7	45	49	53	147
	8	44	44	46	134
	9	41	50	49	140
	10	36	35	41	112
Sums		549 +	521 +	543 =	1,613
Squares		31,689 +	28,027 +	29,993 ≠	267,141
CAB	11	82	64	56	202
	12	79	70	44	193
	13	70	61	46	177
	14	66	44	44	154
	15	62	46	27	135
	16	59	60	38	157
	17	58	48	33	139
	18	56	37	27	120
	19	45	47	24	116
	20	40	26	21	87
Sums		617 +	503 +	360 =	1,480
Squares		39,671 +	26,927 +	14,132 ≠	230,718
BCA	21	91	50	67	208
	22	80	31	55	166
	23	73	57	52	182
	24	70	35	53	158
	25	66	53	44	163
	26	62	37	28	127
	27	58	22	35	115
	28	56	21	44	121
	29	50	26	46	122
	30	47	23	46	116
Sums		653 +	355 +	470 =	1,478
Squares		44,319 +	14,243 +	23,140 ≠	227,812
Overall Sums		1,819 +	1,379 +	1,373 =	4,571
Overall Squares		115,679 +	69,197 +	67,265 ≠	725,671
		Day 1	Day 2	Day 3	
Sums for Days		1,264	1,608	1,699	
Total Sum of Squares		115,679 +	69,197 +	67,265 =	252,141

Procedure for obtaining the sums of squares shown in Table V is:

$$\text{Total} = \frac{1}{90}\left[90(252,141) - (4571)^2\right] = 19,984.9889$$

$$\text{Orders} = \frac{1}{90}\left\{3\left[(1613)^2 + (1480)^2 + (1478)^2\right] - (4571)^2\right\} = 399.0889$$

$$\text{Forms} = \frac{1}{90}\left\{3\left[(1819)^2 + (1379)^2 + (1373)^2\right] - (4571)^2\right\} = 4361.6889$$

$$\text{Days} = \frac{1}{90}\left\{3\left[(1264)^2 + (1608)^2 + (1699)^2\right] - (4571)^2\right\} = 3509.3556$$

Individuals
within orders =

$$\frac{1}{30}\left\{10(725,671) - \left[(1613)^2 + (1480)^2 + (1478)^2\right]\right\} = 9335.2333$$

Forms × days interaction within orders (latin square error) =

$$\frac{1}{90}\left\{9\left[(549)^2 + (521)^2 + (543)^2 + (617)^2 + (503)^2 + (360)^2 + (653)^2 + (355)^2 + (470)^2\right]\right.$$

$$-3\left[(1613)^2 + (1480)^2 + (1478)^2\right] - 3\left[(1819)^2 + (1379)^2 + (1373)^2\right] - 3\left[(1264)^2 + (1608)^2\right.$$

$$\left.\left. + (1699)^2\right] + 2(4571)^2\right\} = 6.1556$$

Residual (individuals × forms interaction within orders) =

$$\frac{1}{30}\left\{30(252,141) - 10(725,671) - 3\left[(549)^2 + (521)^2 + (543)^2 + (617)^2 + (503)^2 + (360)^2\right.\right.$$

$$\left.\left. + 653)^2 + (355)^2 + (470)^2\right] + \left[(1613)^2 + (1480)^2 + (1478)^2\right]\right\} = 2373.4667$$

Of course, either one of the last two terms could be secured more simply by subtraction, but the longer method provides a check upon the accuracy of the computations. Note also that the "latin square error" is the difference between the sum of squares for orders × forms interaction and the "days" effect:

$$\frac{1}{90}\left\{9\left[(549)^2 + (521)^2 + (543)^2 + (617)^2 + (503)^2 + (360)^2 + (653)^2 + (355)^2\right.\right.$$

$$\left. + (470)^2\right] - 3\left[(1613)^2 + (1480)^2 + (1478)^2\right] - 3\left[(1819)^2 + (1379)^2 + (1373)^2\right]$$

$$\left. + (4571)^2\right\} - \frac{1}{90}\left\{3\left[(1264)^2 + (1608)^2 + (1699)^2\right] - (4571)^2\right\} = 3515.5111$$

$$- 3509.3556 = 6.1556$$

Thus this "latin square error" sum of squares is simply the summation over individuals, orders, and forms of the squared residuals of the various order-form means corrected for orders, forms, and days:

$$\sum_i \sum_o \sum_f (\overline{X}_{of} - \overline{X}_o - \overline{X}_f - \overline{X}_d + 2\overline{X})^2.$$

MEANS OF THE THREE FORMS, THREE ORDERS, AND THREE DAYS

		Form				
		A	*B*	*C*		
ORDER	ABC	54.90	52.10	54.30	53.77	= ABC Mean
	CAB	61.70	50.30	36.00	49.33	= CAB Mean
	BCA	65.30	35.50	47.00	49.27	= BCA Mean
		60.63	45.97	45.77	50.79	= Grand Mean

First Day = 42.13 Second Day = 53.60 Third Day = 56.63

It is obvious from the above table that the means for Forms *B* and *C* differ extremely little (0.20 points), while Form *A* is much easier. The appropriate *F*-test for the small difference between 45.97 and 45.77 is:

$$F_{1,54} = \frac{\frac{1}{60}\left\{2\left[(1379)^2 + (1373)^2\right] - (1379 + 1373)^2\right\}}{43.95} = \frac{0.6000}{43.95}$$

The means for the first and second days differ much more than those for the second and third days: 11.47 versus 3.03 points. To test the significance of the difference between 53.60 and 56.63, use the following ratio:

$$F_{1,54} = \frac{\frac{1}{60}\left\{2\left[(1608)^2 + (1699)^2\right] - (1608 + 1699)^2\right\}}{43.95} = \frac{138.02}{43.95} = 3.14 \qquad P > .05$$

If the overall *F*-test for differences among the three order means had been significant, we could have tested the discrepancy between CAB and BCA order means by

$$F_{1,27} = \frac{\frac{1}{60}\left\{2\left[(1480)^2 + (1478)^2\right] - (1480 + 1478)^2\right\}}{345.75} .$$

This presupposes that the differences between tests in their residual effects are negligible. If not, the differences between the order groups will be exaggerated and the significance test vitiated. See Kempthorne (13: Ch. 29).

Completely Permuted 3 × 3 Greco-Latin Square Designs

As noted in the article, there are 36 different rows possible for the greco-latin design with three levels of each of the variables. Using the capital latin letters *A*, *B*, and *C* to represent the latin variable and the arabic numerals 1, 2, and 3 to represent the greek variable, we list the 12 greco-latin squares that these form (the order of rows within each square is immaterial for the type of problem dealt with in this paper):

Square Number	I	II	III	Square Number	I	II	III
1	A1	B2	C3	7	C2	A1	B3
	B3	C1	A2		B1	C3	A2
	C2	A3	B1		A3	B2	C1
2	A1	B3	C2	8	C3	A1	B2
	C3	A2	B1		A2	B3	C1
	B2	C1	A3		B1	C2	A3
3	A1	C2	B3	9	B2	C3	A1
	C3	B1	A2		C1	A2	B3
	B2	A3	C1		A3	B1	C2
4	A1	C3	B2	10	B3	C2	A1
	B3	A2	C1		A2	B1	C3
	C2	B1	A3		C1	A3	B2
5	B2	A1	C3	11	C2	B3	A1
	C1	B3	A2		B1	A2	C3
	A3	C2	B1		A3	C1	B2
6	B3	A1	C2	12	C3	B2	A1
	A2	C3	B1		A2	C1	B3
	C1	B2	A3		B1	A3	C2

By assigning 36 individuals randomly to the 36 rows and analyzing the results without regard to squares, one obtains tests of the three main effects free from confounding with the various untested interactions. Rows and individuals are completely confounded, but this is probably unimportant because the row factor is of no direct interest in test-comparison studies, though it can serve as a check upon the differential residual effects of the treatments (see 13: Ch. 29) if we assign at least two subjects randomly to each of the 36 rows. This design yields the following d. f.: 2 each for roman numerals, latin letters, and arabic numerals; 35 for rows; 36 for individuals within rows; and 72 for individuals × roman numerals interaction within rows, to which would be added the (36-1)(3-1)-2-2=66 that in an incompletely permuted design would be called "square uniqueness" (Grant) or "[greco-] latin square error" (Edwards) or "roman numerals × latin letters × arabic numerals interaction within rows" (Lindquist), giving 138 d. f. for the "within" error term used in testing all of the effects except rows, for which the "individuals within rows" mean square is appropriate.

When working with the 3 × 3 greco-latin square design or modifications thereof, it is helpful to keep in mind that for a single such square there are no *d.f.* for the residual, since $[(3 \times 3) - 1] - 2 - 2 - 2 - 2 = 0$. Thus, as for a 2 × 2 latin square, each effect is completely confounded with the interaction of the other effects. When only 36 individuals are tested, we are in effect using as the "within" error term three pooled interactions, each with 11 × 2 *d.f.*: 12 squares × 3 roman numerals; 12 squares × 3 latin letters; and 12 squares × 3 arabic numerals. There seems to be no reason for thinking that these interactions are not unbiased estimates of experimental error if differences among the treatments in their residual effects are negligible.

Testing 36 boys randomly assigned to each of the 36 possible rows and 36 girls assigned likewise yields a design somewhat similar to the one set forth by Archer (1:524) in his Table I. His scheme calls for three different 4 × 4 squares for the boys and three *other* 4 × 4 ones for the girls, but the method of analysis is identical; see his pp. 523-529, especially p. 527. In the present discussion, an added refinement would be to divide the total sum of squares between the 72 individuals into three components instead of two: between rows, with 35 *d.f.*; between sexes, with 1; and the residual (interaction of rows with sex), with 35. This is possible because each boy has a "mate" among the girls who follows exactly the same order as he; in this sense there are 36 row means for the 72 testees. The residual mean square would be the error term for both rows and sex, but it seems unlikely to the author that this triple partitioning will in actual practice be worth the extra effort.

If 36 boys and 36 girls are tested in each of several school grades, say three different ones, the design will be as shown on the following page.

There are $3(216) - 1 = 647$ *d.f.* Of these, $216 - 1 = 215$ are for differences among the over-all means of the 216 individuals (differences between individuals), and $(216 - 1)(3 - 1) + (3 - 1) = 2(216) = 432$ are for differences "within individuals."

The between-individuals sum of squares may be partitioned into four independent components, with the following *d.f.*, if "rows" are disregarded: grades, 2; sex, 1; grades × sexes interaction, 2; and differences among individuals within the six grade-sex groups, $6(36 - 1) = 210$. If the 36 different greco-latin rows are taken into account, the 210 *d.f.*, would be further subdivided into 35 for rows, 35 for rows × sex interaction, 70 for rows × grades interaction, and 70 for the residual (rows × sexes × grades interaction).

The 13 components of the "within" variance are secured from roman numerals, latin letters, and arabic numerals, their interactions (6 simple and 3 second-order) with grades and/or sex, and the residual within grade-sex groups, which has $6[(36 - 1)(3 - 1) - 2 - 2] = 396$ *d.f.*

The analysis of these designs is jeopardized when significant greco-latin order differences or interactions of orders with other effects occur *because of differences among the treatments with respect to their residual effects* (not just because the individuals differ from row to row). Kempthorne (13) explains why this is so in his Chapter 29, especially on pp. 596-597 and 606-607. The design in Tables VIII and IX of the present writer's article includes two persons for each of the 36 greco-latin orders, thus permitting significance tests of the "order" effect (differential residual effects) and of the three interactions of order with grade and/or sex.

School Grade or Class	Sex	Individual	I	II	III
U	*M*	1	A1	B2	C3
		.	.	.	.
		36	B1	A3	C2
	F	37	A1	B2	C3
		.	.	.	.
		72	B1	A3	C2
V	*M*	73	A1	B2	C3
		.	.	.	.
		108	B1	A3	C2
	F	109	A1	B2	C3
		.	.	.	.
		144	B1	A3	C2
W	*M*	145	A1	B2	C3
		.	.	.	.
		180	B1	A3	C2
	F	181	A1	B2	C3
		.	.	.	.
		216	B1	A3	C2

THE INFLUENCE OF SOCIO-CULTURAL CHARACTERISTICS ON EDUCATIONAL OPPORTUNITIES IN PUBLIC SCHOOL INSTRUMENTAL MUSIC*

MELVIN L. ZACK
Los Angeles State College

Introduction

THROUGH THE education of youth, an attempt is made, in most cultures, to maintain standards and foster ideals considered to be of primary importance. Inasmuch as equality of opportunity has been a cultural ideal of the United States, public education has been expected not only to expound this ideal, but to become the agency through which this ideal could be translated into reality. For the most part, both lay and professional opinion are in agreement with the concept that the abilities and interests of the child should be the only limitations placed upon his participation in any school-sponsored activity.

Many studies, however, have indicated that differences in children's backgrounds due to socio-economic status, color, religion, nationality background, or family size interfere with the attainment of equitable educational opportunities in public schools.[1, 2]** Most of these studies, however, have failed to take into account individual differences in the abilities and interests of children. The present study was designed with the specific purpose of considering two sets of variables in their influence on participation. Not only was the influence of socio-cultural variables, such as color, to be determined, but also the influence of abilities and interests, as well as the interaction between these two sets of variables.

In an investigation concerned with many variables, the validity of the results is determined largely by the adequacy of the design. It is for this reason that the method of procedure is given in considerable detail.

Statement of the Problem

The problem was first to determine the ability-interest differences between participants and non-participants in instrumental music; second, to determine the socio-cultural differences between participants and non-participants; and, most important of all, to take both sets of differences into account in determining whether or not the public schools of Kansas are providing to all children equal opportunities for participation in instrumental music.

The need for this investigation arose from a consideration of the following conditions and assumptions: It has been assumed that instrumental music courses are both elective and selective. They are elective in the sense that no child is required to participate. It has been assumed that any child who wishes to participate may do so. With respect to the selective character of instrumental music courses, it has been assumed that children with appropriate abilities and interests tend to enter and remain in such courses, whereas children who lack these abilities and interests either fail to enter, or enter only to drop out after a period of participation.

Two conditions in Kansas cast suspicion on the validity of these assumptions. One condition is that a majority of schools furnish at public expense only a fraction of the equipment which a student must own in order to participate in band or orchestra. This condition suggests that individual differences in socio-economic status may override individual differences in ability and interest in influencing participation. The second condition is that high prestige is generally associated with membership in a school instrumental group. This condition suggests that individual differences in status due to color, nationality background, religious preference, or size of family may override individual differences in ability and interest in their influence upon participation.

The possible contradictions between these assumptions and conditions, as discussed above, made it necessary to take into account both socio-cultural characteristics as well as appropriate abilities and interests in studying participation in instrumental music classes.

What are the abilities and interests which could be regarded as legitimate requirements for participation in the instrumental music offerings of the public schools? A review of the literature on music education pointed to four factors: musicality, intelligence, the youngster's interest for music, and the support for music in the youngster's home.[3, 4, 5, 6]

---

* A summary of A Study of the Influence of Selected Socio-cultural Characteristics on Participation in Instrumental Music, unpublished doctoral dissertation, University of Kansas, 1953.

** Footnotes will be found at the end of this article.

With the selection of these four variables as the abilities and interests which could be regarded as legitimate limitations upon participation, and with the selection of socio-economic status, color, religion, nationality background, and family size as the socio-cultural characteristics which could not be regarded as legitimate limitations; the problem of the present study became a search for answers to two specific questions:

1. Are participants in instrumental music significantly superior with respect to abilities and interests as compared to non-participants and as compared to drop-outs?
2. When ability and interest differences are taken into account, do individual differences in status due to socio-cultural characteristics have a significant association with participation or with drop-out?

Selection of the Sample

The sample had to include youngsters who had reached that point in school where, if they were not already participating or had not already participated in instrumental organizations, it was very unlikely that any future participation under high school auspices would take place. This criterion simply meant that a part of the sample should consist of seniors in their final semester of high school study.

Because it also was desired to include potential high school drop-outs in the sample, a part of the sample had to include children who were at that level in school immediately prior to a year of high incidence of general drop-out. Since the ninth grade is a year of high general drop-out, it was decided the sample should include eighth grade students.

There still remained the problem of selecting these samples from the total population of eighth grade students and seniors in Kansas schools. Inasmuch as color, religion, nationality background, and socio-economic status were four of the characteristics whose influence upon instrumental participation was being measured, it was desired to select schools from communities where these characteristics apparently were most varied. The rationale for this procedure was obvious. For example, in an all-white community it would be impossible to ascertain the influence of color upon participation. For this reason the most recent census data available on racial structure, nationality structure, educational structure, and occupational structure were utilized[7]. Because census data did not include religious preferences of the population, it was impossible to derive ratios for this characteristic.

Each city in Kansas was ranked in comparison with all other cities in its population bracket with respect to the size of the following ratios:

1. Negro population to total population.
2. Foreign-born white population to native-born white population.
3. Adult college graduates to all adults.
4. Adults with no years of formal schooling to all adults.
5. Employed males in the professional level occupations to all employed males.
6. Employed male laborers to all employed males.

The larger the ratio, the higher was that city's rank. These six ranks were averaged and the cities then were ranked on the basis of this single average. Eight school systems in cities with the highest average ranks were invited to participate in this study. Two of the eight declined. Another school system in a high ranking community, in the same population bracket with the two which had declined, was selected, and its school officials agreed to participate.

This sampling procedure resulted in a total sample of 960 students, of whom 538 were eighth graders and 422 were seniors.

Devices for Securing Data

Intelligence Test Information

In order to help secure the cooperation of schools selected for the study it was decided to administer the Otis Quick-Scoring Tests of Mental Ability, Gamma and Beta, to seniors and eighth graders, respectively, in those school systems which desired this service. In other schools the record of any standard group intelligence test results was utilized. Within each school and at each grade-level, IQ's were converted to *Z*-scores with a mean of 50 and a standard deviation of 10.

This procedure was followed in order that children measured by different tests of intelligence might be compared with each other on the basis of their deviation from the mean of their own group. *Z*-scores were also needed in order to derive a composite score based on the means and standard deviations of the following four measures: intelligence, musicality, interest for music, and home support for music.

Musicality

Measurement of musical ability was obtained by administering A Test of Musicality to all subjects. The items on pages two and three of this test are designed to measure "the excellence of the ability of the individual to apprehend musical structures."[8] According to one authority these are the kinds of items which truly measure musical aptitude or musical ability.[9] This same authority maintains that two widely used tests in this area do not measure musical abilities but rather measure acoustical or sensory abilities.

Interest for Music

Though most experienced music educators would not be opposed to the proposition that a youngster's attitude toward music, or interest for music, is highly indicative of the degree of use to which he puts his musical ability, only one test which measured this important predictive factor could be found. This was the first section, i. e., page one, of A Test of Musicality.

Because norms for this section of the test, as well as for the musicality section, are presented both by grade levels and by sex, it was necessary first to convert all raw scores to the percentile ranks appropriate to the grade placement and sex of the youngsters taking the tests and then to convert these percentile ranks to Z-scores. This procedure made it possible to attain comparability of boys' and girls' scores as well as comparability of eighth graders' and seniors' scores. It also was needed for the composite scores mentioned previously.

Home Support for Music

Hardly any educator would deny that a powerful influence is exerted by the attitudes of parents upon the behavior of their children. In spite of this fact, there was found no instrument which attempted to measure parental attitudes toward music. For this reason an attempt was made to construct such an instrument. Extreme care was taken to insure that the questionnaire measured what it purported to measure, and did so consistently. Two different groups of experienced music educators examined each item in an effort to assess its validity and discriminating power. Statistical estimates of validity were determined by contrasting scores with teachers' dichotomous categorizations of the homes. All t-tests were significant at the .01 level. The reliability of the questionnaire was determined by dividing it into two roughly equivalent halves, scoring each half, and then computing the coefficient of correlation between the scores on these two sections. Inasmuch as four sets of norms had been necessary to handle both sex and grade differences, the correlations were performed on four randomly selected groups, one from within each norm group. The coefficients ranged from .80 to .89, and all were significant at the .01 level.

Socio-cultural Characteristics

Information concerning the occupation of the head of the household, color, religion, nationality background, number of children in the home, and music courses taken was obtained by completion of a questionnaire by the subjects. Wherever school records permitted it, the student report was validated against school data.

The occupational scale developed by Warner, Meeker, and Eells was utilized as a measure of socio-economic status.[10] Protestant sects were categorized in accordance with several articles in a recent encyclopedia of religion.[11] Nationality background was categorized for white subjects only. Grandparents' place of birth was the arbitrary means employed to place a white subject in "native", Anglo-Saxon, Latin, or Slavic categories.

Normality of Data

In order to determine whether the intelligence Z-scores, musicality Z-scores, interest for music Z-scores, and home support Z-scores departed from normality, chi-square tests of significance were applied to these distributions. Since the assumption that these four variables are distributed normally in a school population had been made, it was decided to transform into normalized T-scores those Z-scores which did depart significantly from normality. This procedure was necessitated by subsequent t-tests which are based on such an assumption.

Composite Scores

For each individual in the study it seemed advantageous to obtain a single composite score representing the individual's intelligence, musicality, interest for music, and home support for music. In deriving a composite score, consideration was given first to discriminant function weightings, i. e., weightings for each of these four variables which would maximize the differences between participants and non-participants. However, the use of composite scores based on discriminant function weightings, in further statistical comparisons which hold constant composite score, would bias the outcomes of such comparisons and, therefore, was rejected. The arithmetic mean of each individual's four T-scores was the composite score utilized.

Although each variable could have been considered separately, the interdependence of these four variables in discriminating between participants and non-participants, as shown by prior research, made more valid the use of the composite score. The awkwardness of simultaneously holding constant four different variables was another reason for using the composite.

Statistical Tools Used

At each grade level in each community t-tests were calculated to test the hypothesis that there was no difference between participants and non-participants with regard to means on composite scores. The assumption which had to be satisfied before the t-test could be applied was that there was no difference between standard deviations of the two groups. This assumption of homogeneity of variances was tested by application of the F-test. In the event that this assumption was not satisfied, the Behrens-Fisher or d-test was used in place of the t-test.

At each grade level the procedure explained above was utilized also in a comparison between participants and drop-outs with regard to means on composite scores.

Analysis of variance and covariance is the technique usually used in cases involving comparisons between groups on one variable, e.g., socio-economic status, holding another variable, e.g., composite score, constant. Because distributions of socio-economic status, family size, color, religion, and nationality background could not satisfy the assumptions of normality and continuity, analysis of variance and covariance could not be utilized. Instead, analysis of variation by the method of ranks was used.[12] This technique is applicable to data of a qualitative character and to data which are not normally distributed.

Inasmuch as this statistical tool, analysis of variation by the method of ranks, is more sensitive when individual cell entries are numerically large, pooling of schools was utilized.

For comparisons involving socio-economic status, only those schools with homogeneous distributions of socio-economic categories were pooled. This was necessary because significantly different distributions of socio-economic categories between schools might obscure the association between participation and socio-economic status within either school, should such differing schools be pooled. Chi-square was the statistical tool applied to test the hypothesis that there was no difference in distributions of socio-economic categories between schools. When this hypothesis was accepted, pooling was accomplished.

Table I illustrates the application of analysis of variation by the method of ranks to a group of eighth graders homogeneous with respect to socio-economic status. The results of this test of significance are reported also in Table III, a summary table of similar tests of significance.

In considering the association between color and participation it was decided to equate the Negro students in the study (N=24) with white students of equal composite score, of equal socio-economic status, and of the same grade level. Chi-square was the statistical tool applied to this equated group to test the hypothesis that there was no difference in the proportions of Negro and white participants.

In some schools, t-tests between the means of composite scores of participants and non-participants resulted in findings of no significant differences between these groups. It was not necessary, then, to hold constant composite scores when determining whether or not these two groups differed significantly with regard to religion, color, socio-economic status, family size, or nationality background. In these situations the statistical tool applied was the significance of the difference between percentages.[13]

Findings

Reference to Table II shows that in eleven of the fifteen comparisons made, participants were superior to non-participants with respect to composite score at the .01 level of significance. One comparison was significant at the .02 level; two comparisons were significant at the .05 level; one comparison failed to reach the .05 level of significance.

It was decided to hold constant composite score, representing interests and abilities, in the eleven comparisons significant at the .01 level and in the comparison significant at the .02 level when determining the association between various status characteristics and participation in instrumental music activities. The results of the application of analysis of variation by the method of ranks to these twelve groups, reported in Table III, indicate that, with composite score held constant, there was found no significant association between participation and each of the following characteristics: socio-economic status, religion, family size, and nationality background.

In dealing with the three comparisons which failed to reach the .02 level of significance with respect to differences between participants and non-participants in composite score, composite score was not held constant when determining the association between participation and each of the four characteristics mentioned above. The results of the application of the significance of the difference between percentages to the data in these three schools is reported in Table IV. These results show that for the two senior groups there was found no significant association between participation and the four status characteristics. In the eighth grade group, however, the association between socio-economic status and participation as well as one of the comparisons dealing with the association between nationality background and participation were significant at

TABLE I

RANKS OF PERCENTAGE OF PARTICIPATION IN INSTRUMENTAL MUSIC FOR SPECIFIED LEVELS OF COMPOSITE SCORE AND OF SOCIO-ECONOMIC STATUS AMONG EIGHTH GRADERS IN TOWNS A, C, AND D

Composite score intervals	Ranks based on percentage of participation in instrumental music by socio-economic status						
	VII	VI	V	IV	III	II	I
65+	5	5	1.5	5	5	1.5	5
55-64	7	2	3	5	4.5	1	4.5
45-54	7	4	1	2	5.5	5.5	3
35-44	5	5	2	1	5	5	5
under 35	4	4	4	4	4	4	4
Sum of ranks	28	20	11.5	17	24	17	21.5
Mean rank	5.6	4	2.3	3.4	4.8	3.4	4.3
Deviation	1.6	0.0	-1.7	-.6	.8	-.6	.3

Theoretical mean = $1/2(p+1) = 4.0$
Sum of deviations squared = 6.9

$$\chi_r^2 = \frac{12n}{p(p+1)} (\Sigma d^2)$$

$$\chi_r^2 = 7.4$$

d.f. =6 $.30 > P > .20$

TABLE II

COMPARISONS BETWEEN PARTICIPANTS AND NON-PARTICIPANTS WITH RESPECT TO COMPOSITE SCORES

School	d.f.	F	Mean of participants	Mean of non-participants	t
			Eighth Grade		
A No. 1	24, 46	1.28	53.64	48.72	3.92***
A No. 2	11, 48	1.33	55.58	48.51	3.67***
B	29, 64	1.06	55.07	47.98	4.83***
C	115, 23	1.98*	54.25	47.68	5.8ƒ
D	38, 15	1.13	57.38	47.46	5.55***
E	22, 12	2.28	57.31	45.61	6.62***
F	24, 2	1.16	62.67	50.4	2.64**
G	22, 5	1.8	58.52	53.5	2.39*
			Twelfth Grade		
A	83, 16	1.61	55.88	48.86	3.88***
B	13, 26	1.16	56.79	48.59	3.72***
C	9, 82	1.32	55.7	49.59	2.79***
D	51, 4	3.99	53.8	46.88	2.22*
E	18, 23	1.04	55.96	45.32	6.32***
F	28, 3	1.88	55.5	47.93	1.95
G	14, 33	1.32	55.53	50.32	3.19***

* Significant at the .05 level

** Significant at the .02 level

*** Significant at the .01 level

ƒ Since the assumption of homogeneity of variances was not met, the Behrens-Fisher d-test was used to test the significance of the difference in means. The obtained value (d=5.8) is significant at the .01 level.

TABLE III

ANALYSES OF THE ASSOCIATION BETWEEN PARTICIPATION AND VARIOUS STATUS CHARACTERISTICS WITH ABILITIES AND INTERESTS HELD CONSTANT

Comparison	d.f.	χ_r^2	Probability
Socio-economic status (seniors in towns A and C)	4	2.00	$.80 > P > .70$
Socio-economic status (seniors in towns B, E, and G)	4	4.2	$.50 > P > .30$
Socio-economic status (eighth graders in towns A, C, and D)*	6	7.40	$.30 > P > .20$
Socio-economic status (eighth graders in towns B, E, and F)	6	4.29	$.70 > P > .50$
Family size (seniors in towns A, B, C, E, and G plus all eighth graders except in town G)	4	3.55	$.50 > P > .30$
Religion (seniors in towns A, B, C, E, and G plus all eighth graders except in town G)	5	8.29	$.20 > P > .10$
Nationality background (seniors in towns A, B, C, E, and G plus all eighth graders except in town G)	3	1.16	$.80 > P > .70$

* A detailed presentation of analysis of variation by the method of ranks, dealing with this comparison, is reported in Table I.

the .02 level.

In eight of the fifteen sample schools there were found to be Negro youngsters. In order to determine whether or not these twenty-four Negroes were participating in instrumental music to as great an extent as white youngsters of equal abilities and interests, it was necessary to select from the eight sample schools concerned those white students whose composite scores were of equal value to some Negro student. In order to remove the influence of socio-economic status upon the outcome of a test of significance, it was decided that equivalent socio-economic status would be another criterion for equating the groups. Chi-square was the statistical technique used to ascertain whether or not there was a significant difference between the proportion of participants to non-participants among the Negro youngsters and this proportion in a matched group of ninety-five white youngsters. The data and test of significance are reported in Table V. The results indicate that there was no significant association between color and participation in instrumental music where grade level, composite score, and socio-economic status were controlled.

In the comparisons dealing with differences in mean composite score between eighth grade participants and eighth grade drop-outs as well as mean composite score differences between senior participants and senior drop-outs, the obtained values of t were significant at the .01 level. In each case, participants were superior to drop-outs with respect to the composite of abilities and interests. These findings indicated that composite score would have to be controlled in all of the comparisons involving participants and drop-outs, comparisons dealing with the association between drop-out and each of the following characteristics: socio-economic status, family size, religion, and nationality background. The results of these comparisons, reported in Table VI, indicate that there was found no significant association between drop-out and each of these status characteristics.

Because there were only ten Negro students in the total group of drop-outs and participants, it was deemed impractical to use a statistical test of significance in determining whether or not there was an association between color and drop-out. Inspection of the data, however, seemed to suggest that color was not associated with drop-out.

Interpretations and Implications

Abilities and interests deemed appropriate to participation in instrumental music classes were found to differentiate participants from non-participants in twelve of the fifteen schools and to differentiate participants from drop-outs in all schools. Though this finding constitutes a necessary step in the process of determining whether or not the public schools of Kansas are providing to all children equal opportunities for participation in instrumental music classes, it does not, taken by itself, provide an answer to this problem.

The purpose of ascertaining whether or not participants differed significantly from non-participants and from drop-outs with regard to intelligence, musicality, interest for music, and home support for music, those characteristics of youngsters which may be regarded as justifiable limitations upon participation, was to utilize valid procedures in attacking the crucial question in this study. This question was as follows: do individual differences in status due to color, religion, socio-economic status, nationality background, or family size have a significant association with participation or with drop-out?

In schools in which participants did not differ significantly from non-participants with respect to the composite of abilities and interests, the manner of answering this crucial question was to ascertain whether or not children of differing religious preference, differing socio-economic level, differing color, differing nationality background, and differing size of family were represented in instrumental music organizations in proportion to their numbers in the total school enrollment. This simple procedure could not be used in schools in which participants did show a significant superiority in abilities and interests over non-participants or over drop-outs. In these latter cases, it was necessary to consider different levels of ability and interest, i.e., to hold constant the composite of interests and abilites, in ascertaining whether or not children of varying socio-cultural characteristics were represented in instrumental music organizations in proportion to their numbers in the different ability-interest levels.

The former technique was used in dealing with three of the sample schools; the latter technique, with twelve of the sample schools. With the exception of one eighth grade group, for which judgment must be suspended, the answers to the crucial question in this study were the same.

Individual differences in status due to color, socio-economic level, religion, nationality background, and size of family were found to have no significant association with participation in instrumental music. In all schools, individual differences in these socio-cultural characteristics were found to have no significant association with drop-out from instrumental organizations.

These findings are encouraging to the student of human affairs in that the evidence revealed that an elective and selective area of public

TABLE IV

ANALYSES OF THE ASSOCIATION BETWEEN PARTICIPATION AND VARIOUS STATUS CHARACTERISTICS IN SCHOOLS IN WHICH PARTICIPANTS AND NON-PARTICIPANTS DID NOT DIFFER IN ABILITIES AND INTERESTS

Comparison	X	Probability
Socio-economic status (seniors)	1.74	.08
Socio-economic status (eighth graders)	2.34	.02
Family size (seniors)	.32	.75
Family size (eighth graders)	.48	.63
Religion (seniors)	.82	.41
Religion (eighth graders)	.54	.59
Nationality background (seniors)	.14	.89
Nationality background (eighth graders)		
Native vs. Anglo-Saxon	2.33	.02
Native vs. Slavic	1.39	.16
Slavic vs. Anglo-Saxon	1.12	.26

TABLE V

PARTICIPANTS AND NON-PARTICIPANTS WITHIN A GROUP OF WHITE AND NEGRO YOUNGSTERS EQUATED FOR ABILITIES, INTERESTS, SOCIO-ECONOMIC STATUS, AND GRADE LEVEL

Color	Participants	Non-participants	Totals
Negro	7 (a)	15 (b)	22*
White	14 (c)	81 (d)	95*
Totals	21	96	117

$$\chi_0^2 = \frac{(ad - bc)^2 \ (a+b+c+d)}{(a+b)(c+d)(a+c)(b+d)}$$

$$\chi_0^2 = 3.54$$

d.f. =1 $.10 > P > .05$

---

* Although it was possible to find several white pupils, each of whom could be matched to the same Negro pupil, it was impossible to find any white students who could be a match for two of the Negro youngsters.

TABLE VI

ANALYSES OF THE ASSOCIATION BETWEEN DROP-OUT AND VARIOUS STATUS CHARACTERISTICS WITH ABILITIES AND INTERESTS HELD CONSTANT

Comparison	d.f.	χ^2_r	Probability
Socio-economic status (all seniors)	5	3.68	$.70 > P > .50$
Socio-economic status (eighth graders in towns B, D, and G)	4	5.8	$.30 > P > .20$
Socio-economic status (eighth graders in towns A, C, E, and F)	4	4.4	$.50 > P > .30$
Family size (all eighth graders)	4	1.77	$.80 > P > .70$
Family size (all seniors)	3	.99	$.90 > P > .80$
Religion (all eighth graders plus all seniors)	4	4.31	$.50 > P > .30$
Nationality background (all eighth graders)	2	1.31	$.70 > P > .50$
Nationality background (all seniors)	2	1.36	$.70 > P > .50$

school instruction was providing children with equal opportunities for participation. These results are in diametric contradiction to much of the present published data regarding student participation in public school activities. This fact suggests that additional checks of the validity of this study should be made.

The interests and abilities selected as appropriate to participation in musical activities may not be the best choices possible. Kansas communities may be too homogeneous to test adequately the issues raised in this study. Since the statistical technique of analysis of variation by the method of ranks is more sensitive to large individual cell entries, a larger sampling might be made.

The contradiction between the results of this study and much past research is due also to the philosophical basis and methodology of the present study. Past research has failed, for the most part, to take into account individual differences in the abilities and interests of children for the public school activity being investigated. Having made such an omission, differences in socio-economic status or in some other socio-cultural characteristic between participants and non-participants, when discovered, have been regarded as proof of unequal opportunities for participation. This is the same, it seems, as stating that equality of opportunity in a democracy is the equivalent of identical participation.

According to the philosophical position taken in the present study, such a viewpoint is invalid. In this study, equality of opportunity was interpreted to mean the full development by the school of each individual's abilities and interests without regard to color, religion, and similar socio-cultural characteristics.

Analysis of variation by the method of ranks was well suited to the central problem of this study. By use of this statistical tool, it was possible to determine the rate or percentage of participation in each of several different levels of interests and abilities for each of several categories of socio-economic status, religion, and other characteristics. With data organized in this fashion, it was possible to determine whether interests and abilities were the more influential characteristics associated with participation in music or whether socio-economic status, for example, was more influential.

It is conceivable that the philosophical position and statistical methodology employed in the present study might prove to be useful in areas other than participation in instrumental music.

FOOTNOTES

1. Carl W. Hansen and Paul W. Terry, "Extra Curricular Activities", Encyclopedia of Educational Research, 1950, 424-429.
2. Herman H. Remmers and Nathaniel L. Gage, "Reshaping Educational Policy", Review of Educational Research, XIX (February 1949), 81.
3. E. Thayer Gaston, "A Study of the Trends of Attitudes Toward Music in School Children With a Study of the Methods Used by High School Students in Sight-reading Music", Kansas Studies in Education, II (1941), 53-56.
4. James L. Mursell, The Psychology of Music (New York: W. W. Norton and Company, Inc., 1937), pp. 320-352.
5. Virginia C. Sobieski, "A Study of Instrumental Drop-Outs of the Lawrence (Kansas) Schools", unpublished master's thesis (The University of Kansas, 1951).
6. Jack R. Stephenson, "A Study of Children's Reasons for Participating in School Music", unpublished master's thesis (The University of Kansas, 1948).
7. Characteristics of the Population, Sixteenth Census of the United States: 1940, Vol. 2, Part 3 (Washington, D. C.: Government Printing Office, 1943).
8. E. Thayer Gaston, A Test of Musicality: Manual of Directions (third edition; Kansas City, Missouri: Streep Music Company, Inc., 1950).
9. Mursell, pp. 323-324.
10. Lloyd W. Warner, Marchia Meeker, and Kenneth Eells, Social Class in America (Chicago: Science Research Associates, Inc., 1949), pp. 140-141.
11. Vergilius Ferm, ed., Encyclopedia of Religion (New York: The Philosophical Library, 1945), pp. 54-55; p. 114; pp. 229-30; pp. 340-341; pp. 457-458; p. 585; p. 605.
12. Palmer O. Johnson, Statistical Methods in Research (New York: Prentice-Hall, Inc., 1949), pp. 172-173.
13. Ibid, pp. 80-81.

SIMULTANEOUS ABSOLUTE SCALING FOR SEVERAL GROUPS*

W. A. GIBSON
Center for Advanced Study in the Behavioral Sciences
Stanford, California

THE METHOD of absolute scaling was developed some years ago (ref. 1) as one solution to the problem of a mental unit of measurement. The method defines the psychological scale of measurement to be such that the distributions of ability for several age groups on that scale are normal in form. A check on whether such a definition is possible for a given set of data is made available, and a way of determining the parameters of the resulting scale is provided. The present paper is concerned with an alternative way of determining these same parameters.

The logic of absolute scaling is not difficult. If a scale of measurement exists such that the distributions for several age groups on it are normal, then the location of any point on that scale can be expressed very simply in terms of its position in any one of these distributions. Let the point X_i have normal deviates Z_{i1} in age group 1, Z_{i2} in age group 2, etc. Then

$$M_1 + \sigma_1 Z_{i1} = M_2 + \sigma_2 Z_{i2} = \ldots = M_r + \sigma_r Z_{ir} = \ldots$$

$$\ldots = M_n + \sigma_n Z_{in}, \qquad (1)$$

where M_r and σ_r are the mean and standard deviation of the distribution for age group r, and n is the number of age groups in a given problem. If the point X_i represents the scale position of a raw score, then its Z_{ir} are the normalized standard scores for that raw score in each of the age groups. If X_i is the difficulty value of a test item, then its Z_{ir} are the normal deviates associated with the proportions of the successive age groups who fail that item. Absolute scaling can be applied either to the percentile ranks of raw scores or to item difficulty data.

Equations (1) define a straight line in an n-space in which $Z_{i1}, Z_{i2}, \ldots, Z_{ir}, \ldots, Z_{in}$ are the rectangular coordinates of a point on the line. The equations then state, in effect, that the n-dimensional graph of the normalized standard scores or normal deviates for the n age groups must be linear. Being empirically determined, the plotted points will not lie exactly in a straight line, but they must come close to it before an absolute scale can be fitted with negligible error.

The original absolute scaling method (ref. 1) involves the plots of the normal deviates for all pairs of adjacent age groups. These plots are used to determine the parameters of the absolute scale. Equations (1) are solved explicitly for Z_{i1} in terms of Z_{i2} to give

$$Z_{i1} = \frac{\sigma_2}{\sigma_1} Z_{i2} + \frac{M_2 - M_1}{\sigma_1} . \qquad (2)$$

A straight line is fitted to the plotted points on the Z_{i1} - Z_{i2} graph, and its equation is written in slope-intercept form. The slope then represents the ratio of the two standard deviations, while the intercept gives the difference between the two means in units of σ_1. The process then shifts to the Z_{i2} - Z_{i3} plot, then to the Z_{i3} - Z_{i4} plot, and so on up to the $Z_{i(n-1)}$ -Z_{in} plot. Thus the scale is constructed step-by-step, each new fitted line providing the mean and standard deviation for one more age group. The process requires n-1 line fittings. The method to be described here yields the parameters for the n age groups simultaneously by fitting the line only once in n-space. A disadvantage of the present method is that it is somewhat wasteful of the available data, but this can be overcome, in large measure, in ways that will be discussed later.

For the purpose of fitting a line in n-space it is convenient to borrow some concepts from the n-dimensional vector geometry that has been so useful in multiple-factor analysis (ref. 2). There the factor loadings of tests are the projections of vectors on reference axes. The vectors of factor analysis all radiate out from a common point—the vertex of the vector configuration, and the vertex happens to be located at the origin of the coordinate system. In the present context each raw score or test item may be represented by a vector, the coordinates of whose terminus are the corresponding Z_{ir}. Before we can talk about the vector configuration, however, we must locate its vertex in some relevant way. The configuration will differ, depending upon where the vectors originate, even though the vector termini are fixed.

---

*This research was supported in part by a University of Illinois Research Board Grant and in part by Research Grant M-503 (C) from the United States Public Health Service to the Psychometric Laboratory of the University of North Carolina.

A natural answer here is to locate the vertex on the best-fitting line. In this way the vector configuration approaches unidimensionality.

An established principle of curve fitting in two dimensions is that the best-fitting straight line for a set of points shall pass through the centroid of those points. This principle is readily generalized to n-space. The vertex of the vector configuration can therefore be made a point on the best-fitting line by placing it at the centroid of the vector termini. It is not difficult to show that an additional property of this position of the vertex is that it minimizes the sum of the squared vector lengths.

With the vertex and the centroid identical, it is convenient for subsequent computations to translate the origin of the coordinate system into coincidence with the very same point. Then the best-fitting line will pass through the origin. It is only necessary to locate one more point on that line in order to write its equation. Again it is convenient to borrow some concepts from factor analysis. If the configuration is fairly long and narrow, then most of the vectors will lie in only two octants of the n-space—the one in which the vector projections on all axes are positive, and the one in which all such projections are negative. It is an easy matter to reverse the direction of those vectors that have nothing but negative coordinates. Then the first octant will contain all or nearly all of the vectors. The centroid of their termini can be taken as the second point on the best-fitting line.

The data in Table I can be used to illustrate the computing routine. Table I contains the normal deviates corresponding to the percentile norms for ages 7-11 on the PMA Perception Test (ref. 3). Many entries at the upper left and lower right in the table are left blank because of the instability of normal deviates for very large or very small percentile ranks. Normal deviates for percentile ranks above 97 or below 3 are not recorded in Table I.

A best-fitting line could be determined for all ten age groups simultaneously, but such a procedure would use only the data within the dotted lines in Table I. A better approach is to block off the data into overlapping rectangular sections (blocks A, B, and C in Table I), so as to include as many of the usable entries as possible. In this way most of the normal deviates can be used in the fitting. A line is fitted to each block separately. Then, in what amounts to a repeated application of the same concepts, these three results can be combined to give a single fitted line for the whole table of normal deviates.

Figure 1 shows the graph of columns 1 and 2 in Table I. The fact that this and the other plots are approximately linear indicates that an absolute scale can be fitted to the present data. Also shown in Figure 1 are the results of the fitting procedure that is being used. The two parallel slanted lines indicate that the points between them are represented in one of the blocks. Here it is block A. The other constructions will be explained as they are made possible by the calculations.

At the bottom of Table I are recorded the sums and means of the column entries within the three rectangular sections. The column means, m_r, for each block are the coordinates of the centroid of the points represented in that block.

In Figure 1 the dotted lines labeled a_{i1} and a_{i2} have been made to intersect in the centroid of the points represented in block A. The projections of all points on axes translated to this new position are obtained simply by subtracting the mean of each column from every entry in the column. That is,

$$a_{ir} = Z_{ir} - m_r . \qquad (3)$$

The resulting a_{ir} for block A are shown in Table II. From that table it is apparent that, with one exception, the projections of all points on the new axes are either all positive or all negative.

In this translation of axes it is convenient to carry along certain points not represented within a particular block, and to determine their position in the new coordinate system. These are the points whose coordinates on some of the axes represented in the block are missing. Section 1 of Table III shows the a_{ir} for such points on the new axes of Table II. These coordinates are obtained by means of equation (3) from the corresponding entries in columns 1-5 of Table I.

It is desirable to relate two other sets of points to the same new axes. These are the points having Z_{ir} of zero and of 1--the mean, M_r, and the point $(M_r + \sigma_r)$ in the distribution of ability for each age group. Imagine Table I as being extended at the bottom so as to list the Z_{ir} for these points in the appropriate columns. Then the coordinates of these points with respect to the new axes can be obtained in exactly the same way as were the entries in Table II and in section 1 of Table III. The results are shown in section 1 of Table IV. In symbols,

$$a_{Mr} = 0 - m_r = - m_r , \qquad (4)$$

and

$$a_{\sigma r} = 1 - m_r . \qquad (5)$$

At the bottom of Tables II, III, and IV are shown summary checks for the a_{ir} contained in those tables. These checks are made possible by summing equation (3) over the q values

TABLE I

NORMAL DEVIATES FOR PMA PERCEPTION TEST

	Age Group									
	1	2	3	4	5	6	7	8	9	10
	Age Range									
Raw Score	7-7.5	7.5-8	8-8.5	8.5-9	9-9.5	9.5-10	10-10.5	10.5-11	11-11.5	11.5-12
43								C		1.64
41								1.88	1.75	1.55
39								1.64	1.64	1.41
37					B	1.88	1.88	1.48	1.41	1.23
35					1.88	1.75	1.55	1.34	1.17	.95
33					1.64	1.48	1.23	1.08	.95	.74
31				1.75	1.55	1.23	1.04	.84	.67	.52
29	A			1.64	1.28	.99	.77	.58	.39	.39
27	1.88	1.75	1.75	1.28	.95	.74	.47	.33	.10	.03
25	1.75	1.55	1.41	.99	.67	.39	.18	.03	-.18	-.28
23	1.48	1.28	1.04	.71	.36	.05	-.10	-.31	-.44	-.55
21	1.17	.99	.67	.50	.05	-.18	-.36	-.55	-.74	-.81
19	.92	.74	.33	.20	-.25	-.47	-.67	-.84	-1.04	-1.04
17	.67	.47	.08	-.10	-.52	-.74	-.99	-1.13	-1.28	-1.28
15	.33	.13	-.25	-.41	-.81	-1.08	-1.34	-1.41	-1.55	-1.55
13	-.05	-.15	-.61	-.74	-1.04	-1.34	-1.64	-1.64	-1.88	-1.88
11	-.47	-.50	-.88	-1.04	-1.28	-1.55	-1.88	-1.88		C
9	-.77	-.81	-1.08	-1.34	-1.48	-1.75	-1.88	B		
7	-1.04	-1.08	-1.28	-1.55	-1.64	-1.88				
5	-1.34	-1.34	-1.55	-1.75	-1.75					
3	-1.64	-1.75	-1.75	-1.88	-1.88					
1	-1.88	-1.88			A					
ΣZ_{rC}								3.32	.97	-.57
m_{rC}								.22	.06	-.04
ΣZ_{rB}					4.48	1.27	-1.74	-3.56		
m_{rB}					.34	.10	-.13	-.27		
ΣZ_{rA}	2.89	1.28	-2.12	-5.13	-8.62					
m_{rA}	.22	.10	-.16	-.39	-.66					

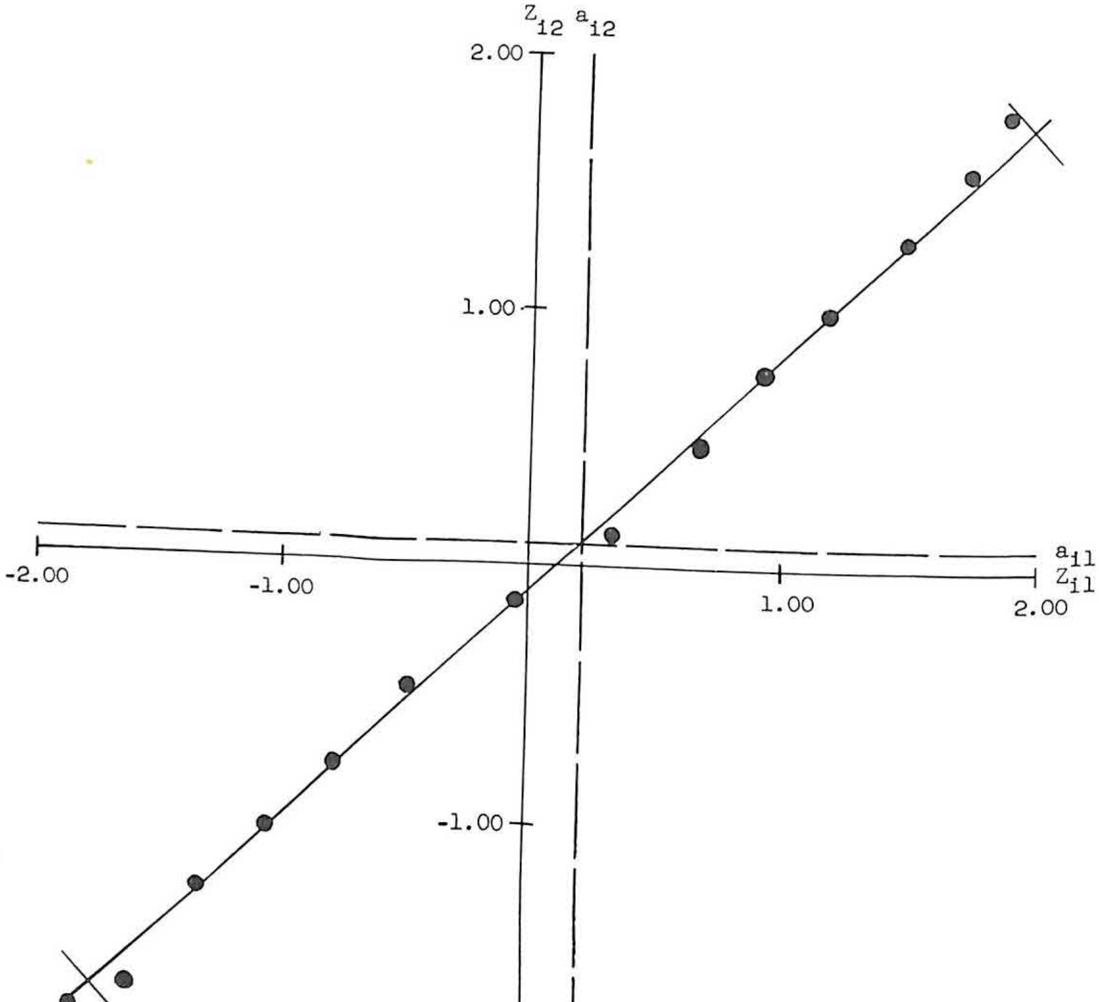


Figure 1

Graph of Normal Deviates in Columns 1 and 2 of Table I

TABLE II

TRANSLATED COORDINATES FOR SECTION A OF TABLE I

Raw Score	Age Group 1	2	3	4	5	V
27	1.66	1.65	1.91	1.67	1.61	102.6
25	1.53	1.45	1.57	1.38	1.33	87.8
23	1.26	1.18	1.20	1.10	1.02	69.8
21	.95	.89	.83	.89	.71	51.8
19	.70	.64	.49	.59	.41	34.5
17	.45	.37	.24	.29	.14	18.4
15	.11	.03	- .09	- .02	- .15	- 1.1
13	- .27	- .25	- .45	- .35	- .38	- 20.3
11	- .69	- .60	- .72	- .65	- .62	- 39.6
9	- .99	- .91	- .92	- .95	- .82	- 55.5
7	- 1.26	- 1.18	- 1.12	- 1.16	- .98	- 69.1
5	- 1.56	- 1.44	- 1.39	- 1.36	- 1.09	- 83.1
3	- 1.86	- 1.85	- 1.59	- 1.49	- 1.22	- 97.5
$\Sigma +$	6.55	6.18	6.24	5.92	5.22	
$\Sigma -$	- 6.63	- 6.23	- 6.19	- 5.96	- 5.11	
Σ	.03	- .02	- .04	- .06	- .04	- 1.3
Ch.	.03	- .02	- .04	- .06	- .04	- 1.5
$b = \Sigma II$	13.18	12.41	12.43	11.88	10.33	730.07 = Σb^2
k_r	55.39	58.83	58.73	61.45	70.67	

TABLE III

TRANSLATED COORDINATES FOR COLUMNS 1-5 OF TABLE I, EXCLUDING SECTION A

	Section 1					Section 2						
	Age Group					Age Group						
Raw Score	1	2	3	4	5	1	2	3	4	5	Σ	V
35					2.54					179.5		179.5
33					2.30					162.5		162.5
31				2.14	2.21				131.5	156.2	287.7	143.9
29				2.03	1.94				124.7	137.1	261.8	130.9
1	-2.10	-1.98				-116.3	-116.5				-232.8	-116.4
Σ	-2.10	-1.98		4.17	8.99	-116.3	-116.5		256.2	635.3		
Ch.	-2.10	-1.98		4.17	8.99	-116.3	-116.5		256.2	635.3		

TABLE IV

TRANSLATED COORDINATES FOR $M_r + \sigma_r$ AND M_r OF COLUMNS 1-5 OF TABLE I

	Section 1					Section 2					
	Age Group					Age Group					
Point	1	2	3	4	5	1	2	3	4	5	V
$M_5 + \sigma_5$					1.66					117.3	117.3
$M_4 + \sigma_4$				1.39					85.4		85.4
$M_3 + \sigma_3$			1.16					68.1			68.1
$M_2 + \sigma_2$		.90					52.9				52.9
$M_1 + \sigma_1$	.78					43.2					43.2
M_5					.66					46.6	46.6
M_4				.39					24.0		24.0
M_3			.16					9.4			9.4
M_2		-.10					- 5.9				- 5.9
M_1	-.22					-12.2					- 12.2
Σ	.56	.80	1.32	1.78	2.32	31.0	47.0	77.5	109.4	163.9	
Ch.	.56	.80	1.32	1.78	2.32	31.0	47.1	77.5	109.4	164.0	

TABLE V

ABSOLUTE SCALE VALUES FOR RAW SCORES

Raw Score	Scale Value	Raw Score	Scale Value
43	4.00	21	1.15
41	3.85	19	.84
39	3.63	17	.54
37	3.46	15	.21
35	3.33	13	- .12
33	3.04	11	- .41
31	2.75	9	- .63
29	2.50	7	- .85
27	2.09	5	-1.04
25	1.79	3	-1.24
23	1.46	1	-1.88

TABLE VI

ABSOLUTE SCALE VALUES FOR M AND σ

Age Group	Age Range	M	σ
10	11.5 - 12	2.07	1.18
9	11 - 11.5	1.96	1.11
8	10.5 - 11	1.76	1.11
7	10 - 10.5	1.58	1.05
6	9.5 - 10	1.34	1.07
5	9 - 9.5	1.07	1.18
4	8.5 - 9	.65	1.11
3	8 - 8.5	.39	1.06
2	7.5 - 8	.11	1.06
1	7 - 7.5	.00	1.00

of a_{ir} that are involved for any value of r. Thus

$$\sum^{q} a_{ir} = \sum^{q} Z_{ir} - qm_r. \qquad (6)$$

For example, the sum of the entries in column 1 of section 1 in Table IV is .56, and by equation (6) this should agree with the check value $(1 + 0 - 2 \cdot .22)$ if the a_{ir} have been correctly calculated. The sum row in Table II is the sum of rows (15), ($\Sigma+$), and ($\Sigma-$) in that table, the latter two rows being used in subsequent calculations.

With all points referred to new axes whose intersection defines one point on the best-fitting line, the next task is to locate a second point on that line. At the bottom of Table II are shown the sums of the positive ($\Sigma+$) and of the negative ($\Sigma-$) coordinates for block A. The point (15) is excluded from the formation of these sums. The absolute sums for these two rows appear in the row marked ($\Sigma||$). These absolute sums are proportional to the coordinates of the point that is wanted. That point is the centroid of the termini of all vectors that lie in the first octant after the reflection of those vectors having all negative coordinates. In the present example, 6 vectors alread lie in the first octant, while 6 others can be put there by mere reversal of their direction. Row ($\Sigma||$) gives the sums of the coordinates of these 12 vector termini on each of the coordinate axes. Division of these sums by 12 would give the coordinates of the centroid being sought. Just as they stand, however, these sums constitute direction numbers for the best-fitting line, and they have been so used to determine its slope in Figure 1. They may be interpreted as the coordinates of a vector colinear with the centroid vector, but 12 times as long. For convenience this long vector will be called the line vector.

The two ends of the line vector have been used to define the best-fitting line. The next task is to determine the projections of all points on that line. Since these points represent vector termini, it is convenient to construe this task as being that of computing the scalar products of all vectors with the line vector. Such scalar products are proportional to the projections of the vectors on the best-fitting line. The proportionality constant is the length of the line vector. Since the scale of measurement along the best-fitting line is essentially arbitrary, the scalar products themselves are sufficient for indicating relative spacings along that line.

If the main body of Table II is called the matrix A, column V is made into a column vector V, and row ($\Sigma||$) is transposed into a column vector B, then the matrix product,

$$AB = V, \qquad (7)$$

gives the scalar products of all vectors in block A with the line vector. In ordinary algebra, the i-th scalar product is the i-th entry in column V and is give by

$$v_i = a_{i1}b_1 + a_{i2}b_2 + \ldots . + a_{ir}b_r + \ldots . + a_{in}b_n, \qquad (8)$$

where b_r is the r-th entry in row ($\Sigma||$). Summing over i in equation (8) gives

$$\sum_i v_i = b_1 \sum_i a_{i1} + b_2 \sum_i a_{i2} + \ldots . + b_r \sum_i a_{ir} + \ldots . + b_n \sum_i a_{in}. \qquad (9)$$

This provides a summational check on the calculation of column V. The quantity on the right in equation (9) is shown in the (Ch.) position at the bottom of column V. Its agreement with the sum of column V constitutes the check.

If point i lies on the best-fitting line and its projection on any axis r is known, the scalar product of its vector with the line vector can be determined. Suppose that b_r has to be multiplied by a quantity c_i in order to give a_{ir}. Then for all axes, including r,

$$a_{i1} = c_i b_1, a_{i2} = c_i b_2, \ldots . , a_{ir} = c_i b_r, \ldots . , a_{in} = c_i b_n. \qquad (10)$$

From equations (8) and (10) the scalar product of vector i with the line vector is

$$v_i = c_i b_1^2 + \ldots . + c_i b_r^2 + \ldots . + c_i b_n^2 = c_i \sum_{r=1}^{n} b_r^2. \qquad (11)$$

The quantity c_i is defined by equation (10) as

$$c_i = \frac{a_{ir}}{b_r}, \qquad (12)$$

so that equation (11) becomes

$$v_i = \frac{\sum_{r=1}^{n} b_r^2}{b_r} a_{ir}. \qquad (13)$$

Let

$$k_r = \frac{\sum_{r=1}^{n} b_r^2}{b_r} \qquad (14)$$

Then

$$v_i = k_r a_{ir}. \qquad (15)$$

The quantity k_r is constant for all values of i.

The quantity Σb^2 is shown at the right of row (Σ II) in Table II. In row (k_r) at the bottom of Table II are recorded the values of k_r. In section 2 of Tables III and IV are recorded the results of multiplying the projections in section 1 by the appropriate k_r. At the bottom of section 2 in Tables III and IV are shown the usual summary checks, based upon the summation of equation (15) over the q values of a_{ir} that are involved for any value of r. Thus,

$$\sum^{q} v_i = k_r \sum^{q} a_{ir} . \tag{16}$$

The entries in section 2 of Tables III and IV are estimates of the scalar products of the corresponding vectors with the line vector, based on the assumption that all of the points lie on the best-fitting line. Where only one estimate is available, it has been recopied into the column marked V in Tables III and IV. When two or more estimates are available for a given point, their sum is recorded in the sum column and their average is recorded in column V. Summary checks are not possible here, but the regular progression of the entries in column V, especially as between Tables II and III, provides an inspectional check, and of course each entry in column V of Table III must be similar in magnitude to the quantities of which it is the average.

The entries in column V of Tables II, III, and IV are a fitted unidimensional representation of the first five columns of Table I. Similar calculations based on sections B and C of Table I give two other fitted dimensions, Hence in one computational cycle the original ten dimensions are reduced to three. These three in turn can be collapsed into one by a single repetition of the same kind of calculations. Nothing essentially new is involved in such calculations, and so they will not be shown here. The final absolute scale is shown in Tables V and VI. The origin and unit of measurement have been adjusted so that age group 1 has a mean of 0 and a standard deviation of 1. The standard deviations in Table VI were obtained from the formula,

$$\sigma_r = (M_r + \sigma_r) - M_r , \tag{17}$$

applied to the adjusted scale.

The computing procedures that have been described here are independent of the graphs and could be completed without them. It is nevertheless a valuable addition to have some of the graphs (for all pairs of adjacent columns, for example) as visual checks. In the first place, the graphs indicate whether an absolute scale is possible at all. If deviations from linearity are serious and cannot be traced to plotting or clerical errors, then the computing process need never begin. In making the constructions it is a frequent experience to find something wrong with them on all of the graphs involving a particular column of normal deviates. This usually indicates that a computing error has occurred somewhere in that column. Furthermore, the nature of the graphical disparity usually suggests the kind of computing error that has been made. The incorrect translation of axes is indicated by the crossing of the new axes away from the middle of the configuration, and errors in computing the direction numbers show up in the tilt of the fitted lines. Computing errors that are actually very small can be detected from the graphs. Finally, the fit of the lines is a visual index of how well the theoretical scale of measurement is reflected in empirical fact.

REFERENCES

1. Thurstone, L. L. "A Method of Scaling Psychological and Educational Tests," Journal of Educational Psychology, XVI (1925), pp. 433-451.

2. Thurstone, L. L. Multiple-Factor Analysis (Chicago: University of Chicago Press, 1947).

3. Thurstone, L. L., and Thurstone, Thelma Gwinn. Examiner Manual for the SRA Primary Mental Abilities, Elementary (Ages 7-11) (Chicago: Science Research Associates, 1948).

AN EXPERIMENTAL EVALUATION OF TWO METHODS OF TEACHING MUSIC IN THE FOURTH AND FIFTH GRADES

CARL B. NELSON
University of Minnesota

WHILE MANY judgments have been made by music educators as to the comparative value of instrumental music training versus vocal music training, very little systematic research has been done in classroom experimentation to obtain evidence of the effects of these judgments. The purpose of this experiment was to obtain objective data with respect to the relative worth of a course of study which utilized not only vocal music training but also instrumental music instruction.

The stated aim of many music teachers and supervisors has been "music for every child". It is of significance to note that surveys conducted in recent years reveal that the majority of children in our schools receive training only in vocal music. These surveys consistently show that instrumental music education is reserved for children who demonstrate outstanding ability in music or for those who can afford to purchase instruments and lessons. It also is apparent that public school instrumental training programs in a majority of cases are sponsored so that bands or orchestras can be developed, particularly at the high school level. Thus, in the latter situation, this type of musical training is not primarily adapted to meet the needs of the individuals but rather pointed toward the development of a performing group.

Although most children have some sort of vicarious instrumental music experience in public schools, it is generally not accepted that a background in instrumental performance is necessary for every child. It seems that music teachers universally assume that instrumental music education is more demanding on pupils so that those who show a greater talent for music are chosen for this training.

Of course there are obvious reasons why such an attitude would be held frequently by music educators. It is considered prohibitively expensive to provide an opportunity for every child to have both instrumental and vocal music instruction. In addition, it would be necessary to obtain staff members to teach in both areas. The item of additional expense for the latter is not the only consideration; music educators also believe that it would be difficult to recruit personnel who would be qualified to handle this sort of teaching. These are major items which appear to stand in the way of a practical approach to a combined vocal-instrumental music course of study.

It was not the purpose of this study to provide possible solutions to these problems. While the writer realizes that they are not easily overcome, the issue investigated is: does this type of program fill a need in our present elementary music course of study? It would seem logical to show first that such a system of instruction is worth the effort. Therefore, this study was undertaken to secure reliable objective information concerning the relative value of a music course including vocal and instrumental participation by the pupils as compared to one including only vocal instruction.

The Population Sample

The population selected for the study was the fourth and fifth grades of the Concord School, Edina, Minnesota. The Concord School is a six-year public elementary school with an enrollment of approximately 500 pupils.

The sample was intact since all the fourth and fifth grade children of this school were included in the experiment. However, the writer found that this particular group of pupils could be considered a representative sample of fourth and fifth grade children in the Edina elementary school system. Random samples of children who had been enrolled in the fourth and fifth grade two, four, and six years previously were compared to the present sample on the basis of intelligence (as measured by the Otis Quick-Scoring Mental Ability Test[1]* and grade point average. The technique of analysis of variance was applied to the data. Conclusions based on the findings of this experiment thus must apply only to fourth and fifth grade children of this school system.

In the previous spring the children had been grouped into four classes by the elementary school principal who was not at liberty to change the grouping according to a preferred random assignment in the fall. That is, the entire sample already had been sorted into four distinct classes. The principal stated that no systematic order was used in the assignation of the children to

* Footnotes will be found at end of this article.

their respective groups. Therefore, interpretations of the finding of this study must recognize this particular manner of selection. Although the precision of the experiment was not impaired by this circumstance, one cannot know for sure whether the error introduced by this selection is an over-estimate or an under-estimate. So that the four subclasses might be equally represented, the records of six children were deleted by random selection. Thus there were, for purposes of analysis, twenty-six children in each of the four subgroups. The total N equaled 104.

A study of the occupations of the fathers of the subjects showed that the group could be regarded as a select sample in a socio-economic sense. The occupation of each parent was compared to the categories listed in the Minnesota Occupational Scale[2]. However, the tabulation showed that no difference existed between the control and experimental groups with respect to that criterion.

The subgroups were likewise canvassed to discover if any differences existed with regard to previous instrumental music training. A tabulation of the number of children who had played instruments prior to the time of the final testing and the number of months of this instruction revealed no differences between the subclasses.

The tests used in this experiment to measure changes in musical behavior of the subjects are an integral part of the experimental design. For this reason the writer calls attention to the selection and construction of these devices.

Measuring Instruments Used

A major problem facing the investigator was to obtain tests to measure achievement in music skills, specifically those connected with reading and understanding music notation. Existing music achievement tests were considered unsuitable for various reasons. The tests to be used had to meet certain specifications in order that results obtained from them would merit justifiable conclusions.

The instruments used should not give biased results because of item content nor the manner of their construction. It would seem extremely important to use tests which would not give unfair advantage to students trained in instrumental music. Tests utilizing content at the fourth and fifth grade level primarily should also be a more valid test of the subjects' learned skills.

The author decided to construct an achievement test of music reading skills to overcome difficulties inherent in those already on the market. The test was divided in two parts called Test 1 and Test 2. Test 1 was designed specifically to measure the pupils' ability to recall notation and terminology used at the fourth and fifth grade level. It also measures the subjects' ability to apply this knowledge to new situations. Test 2, while it demands the same information as Test 1, was constructed to measure the childrens' ability to match audio stimuli to visual stimuli.

Courses of study in music at the elementary level filed at the curriculum laboratory of the University of Minnesota were studied and materials and objectives of the music curriculum at the Edina elementary school were examined. These sources aided the author to determine the scope of the test. Preliminary forms of the two tests were drawn up and submitted to three experts in the field of music education, one of whom was the elementary music supervisor of the Edina system. It was through these steps that the investigator established the content validity of the instruments.

Estimates of the reliability of both Test 1 and Test 2 were made utilizing Hoyt's technique by means of the analysis of variance[3]. Reliability coefficients on the final testing for Test 1 ranged from .89 to .95 in the subgroups. For Test 2, the coefficients ranged from .72 to .81. Both tests were found to measure sufficiently accurately to differentiate among individuals.

It would be of considerable value to discover if the effect of instrumental music instruction influenced the subjects' preference for various categories of music. To test this, the author selected the Keston Music Preference Test[4] This instrument was devised by Professor Morton J. Keston of the Department of Psychology at the University of New Mexico as a means of measuring musical preferences of senior high school students in a study he conducted in 1946-1947.

The test is composed of thirty items; each item consists of excerpts from four orchestral compositions. Because each of the excerpts is approximately 45 seconds in length, all four fit on one side of a 12-inch disc. The groupings are arranged so that the same mood is expressed in each of the four excerpts. Due to this arrangement, it is possible to say that one is not attempting to measure the preferences of the subjects with respect to mood.

The subject is instructed to rank the four excerpts of each item in order of his preference. The four categories are (a) severely classical, (b) serious popular classical, (c) easy popular classical, (d) popular.

Keston established the validity of his test by administering it to twelve expert musicians and to students rated by their teachers as musical. In addition, he compared his categories of excerpts with lists of the selections in phonograph record catalogues.

Circumstances in the present study did not permit an estimate of the reliability of the test by means of a test-re-test, which would appear to be preferable. Keston established that, with his sample, the test was highly reliable. In the present study, a split-halves test of reliability showed coefficients ranging from .88 to .97 in the subclasses (final testing). Keston selected a portion of his sample to discover their consistency in making preferences on this test. He presented the four excerpts of each item in the six possible paired comparisons. By means of the "circular triad"[5] criterion, he was able to demonstrate a high degree of consistency in their choices.

Other Measures

I.Q. - This measure was obtained from the pupils' individual school record folders; as mentioned earlier, the Otis Test (Alpha) was the instrument used. In most cases, the tests had been administered to the children when they were in the third grade.

Grade point average - To obtain a measure of school achievement, the writer procured the office record cards for each child and calculated the grade point average for each. Only the letter grades earned during the year of this study were used.

Reading comprehension - These scores were secured from the classroom teachers' records immediately after the spring testing program. The reading comprehension test is a subtest of the Iowa Every-Pupil Test of Basic Skills, Elementary Battery, Form L[6].

Procedure of the Study

Both the control and experimental subgroups held music class daily for the entire school year. In order to minimize the effect of the instructor, the investigator taught all four subgroups from September 3, 1952, when school began in the fall, until May 22, 1953, the close of the school year. Because of the physical environment in which the experiment was staged it was not possible to differentiate the control and experimental treatments until November 17. Thus the period under investigation extended actually for twenty-five weeks. The class period in all groups was one-half hour long.

The control subgroups followed the course of study in general use throughout the elementary system. The great majority of time was used in singing with a good deal of emphasis placed on reading musical notation. The solfeggio system and the movable "do" were used in sight singing almost exclusively.

The only difference between the experimental and control groups' course of study was the introduction of instrumental instruction. During the 25-week period, exactly half of the class time devoted to music was spent in this manner. The division of time for the experimental treatment was arranged so that alternate days were spent for instrumental and vocal music training. The unit of organization in both types of instruction was the entire class.

Materials for the singing classes in both the experimental and control groups were furnished by the school. The American Singer series[7] and the World of Music series[8] supplied the author with the primary source of song material. Nearly all of the materials for the instrumental program were furnished by the writer. Instruments were rented for the childrens' use from a local music store. The investigator also furnished the children racks, reeds, instrument repair material and method books. The method book primarily used was The Boosey and Hawkes Band Method[9].

The Design of the Experiment

In order to obtain as much information as possible in the natural setting of the public school, the author decided to use a 2 X 2 factorial design. Not only would it be of value to study the effect of the experimental treatment, but also the effect of the treatment at different levels. Thus, by establishing control and experimental groups at both the fourth and fifth grade levels, the study yielded more pertinent information. The investigator felt that the efficiency of the experiment was correspondingly increased and the results more widely applicable.

The assignment of the instructional method to the two groups at each level was decided randomly. Each class was numbered either "one" or "two". By drawing a number from a table of random numbers, it was decided whether the even- or odd-numbered class should receive the experimental treatment.

The measuring instruments were applied at the beginning of the experiment and again at the end to determine what differences might have been effected by the introduction of instrumental instruction. The analysis of variance and covariance provided the means of testing the significance of the differences.

Analysis of the Data

The analysis of variance and covariance was a useful test in this particular experiment because it revealed the significance of a number of components simultaneously. Furthermore, this tool increased the precision of the analysis by making adjustments in the outcomes which may have been due to the inequalities in initial abilities or achievement.

Choice of independent variates - Data on six variables had been collected, namely, the pretests of the two music achievement tests and the Keston Test, I.Q., grade point average and reading comprehension. It seemed desirable to determine whether or not intellectual capacity, achievement and reading comprehension showed significant relationships with the criteria. Clearly, it would be undesirable to assume that such factors were unrelated to the criteria without first investigating the relationships.

All the correlations between each of the independent variates (reading, school achievement, I.Q. and the three pretests) and the correlations between them and the dependent variates (the criterion measures) were computed. Thus, for each of the dependent variates, Test 1 final, Test 2 final and the final test of the Keston, six systems of simultaneous equations were solved for the four subgroups. The method used in this study to solve for the unknown quantities was the well-known Doolittle method[10].

The standard partial regression coefficient (beta) for each variate was then computed for each subgroup on each of the dependent variates. To test the significance of the beta's it was necessary to obtain the ratio value of the beta to its standard error in each case. Since these values were distributed as "t", the significance of the beta's was determined by referring to the "t" table with nineteen degrees of freedom.

The standard partial regression coefficients of the pretests of the three dependent variates were the only significant contributors to the criteria. No significant relationship with a criterion variable was found for intelligence, achievement, or reading comprehension in any subgroup at the one percent level.

The next step in the analysis consisted of the application of the analysis of variance and covariance to test the hypothesis that there was no difference between the experimental and control groups on each of the three criterion variables. Since the preceding analysis revealed that the pretests were significantly related to the criteria, only these variables were controlled by covariance.

The assumption of equal variance on all three of the tests for the four subgroups was met by applying the L_1 criterion[11]. The method used to test the assumption of a normal distribution on each of the measures was the probit test[12]. It was found that there were no significant deviations from that distribution. Table I presents the means and standard deviations on the pretests and final tests of the four subclasses for all three variables.

Knowledge of musical notation - Table II gives the analysis of variance and covariance table, with respect to Test 1 data, for interaction between method and grade. Since the experimental treatment was introduced at both the fourth and fifth grade levels, this interaction effect must be tested first. Note that there is a significant interaction. Due to this effect, the variation attributed to the interaction was not pooled with the residual error in testing the null hypothesis that there is no significant difference in the means.

The adjustment of the sums of squares to test the hypothesis of no difference between the means of the experimental and control groups on Test 1 is indicated in Table III. The tabled value of F at the one percent level when $n_1 = 1$ and $n_2 = 99$ is 6.90. The value of F obtained for "between methods" in this case is 27.38. Thus the experimental classes achieved significantly higher than the controls with respect to skills measured by Test 1.

Because it was shown that a significant interaction existed between the main effects of treatment and level, it was of interest to discover whether or not there was a significant difference in adjusted means for individual pairs of subclasses. A method of testing this is described by Lindquist[13] and Mood[14].

The criterion means for each of the groups were adjusted by the following formula:

$$\bar{Y}_i' = \bar{Y}_i - b_w(\bar{X}_i - M_x)$$

where

$\bar{Y}_i'$ (i = 1, 2) is the adjusted criterion mean for the ith group.

$\bar{Y}_i$ (i = 1, 2) is the criterion mean for the ith group.

b_w is the "within subclasses" regression coefficient.

$\bar{X}_i$ (i = 1, 2) is the mean of the control variable for the ith group.

M_x is the grand mean of the control variable.

The error variance of the difference between the two adjusted criterion means is given by

$$MS_w' \left[\frac{1}{n_1} + \frac{1}{n_2} + \frac{(\bar{X}_1 - \bar{X}_2)^2}{SS_{w(x)}}\right]$$

where

MS_w' is the adjusted mean square for "within subclasses".

$SS_{w(x)}$ is the sum of squares for "within subclasses for the control variable.

n_1 is the number of individuals in the first group.

n_2 is the number of individuals in the second group.

The value of "t" therefore, is equal to

TABLE I

MEANS AND STANDARD DEVIATIONS ON THE CRITERION VARIABLES BY SUBGROUPS

Subgroup	Pretest		Final	
	Test 1			
	$\bar{X}$	S. D.	$\bar{X}$	S. D.
Fourth grade control	23.69	8.86	26.15	9.87
Fourth grade experimental	20.73	10.58	27.27	12.05
Fifth grade control	23.83	8.96	22.96	9.12
Fifth grade experimental	24.34	9.67	34.50	8.44
	Test 2			
	$\bar{X}$	S. D.	$\bar{X}$	S. D.
Fourth grade control	17.46	12.62	24.27	11.26
Fourth grade experimental	19.69	9.97	24.81	10.37
Fifth grade control	19.23	9.29	20.12	12.97
Fifth grade experimental	19.12	12.21	26.85	11.45
	Keston Test *			
	$\bar{X}$	S. D.	$\bar{X}$	S. D.
Fourth grade control	99.04	24.77	104.50	27.70
Fourth grade experimental	110.77	17.99	103.54	24.77
Fifth grade control	108.04	21.80	116.15	30.74
Fifth grade experimental	95.15	26.72	89.15	29.93

* "Poor" performance on this test is indicated by a high score

TABLE II

ANALYSIS OF VARIANCE AND COVARIANCE TO TEST THE HYPOTHESIS OF NO INTERACTION BETWEEN GRADE AND METHOD WITH RESPECT TO TEST 1 FINAL SCORES

Source of Variation	d. f.	Σy^2	Σx^2	Σxy	Adjusted d. f.	 S. S.	 M. S.	 F	 Hyp.
Within subclasses	100	9,927.9616	9,106.6924	6,889.0	99	4,716.5941	47.64		
Interaction	1	706.1635	100.0384	265.7885	1	361.2847	361.28	7.58	Reject

TABLE III

ANALYSIS OF VARIANCE AND COVARIANCE OF TEST 1 FINAL SCORES WITH TEST 1 PRETEST SCORES HELD CONSTANT

Source of Variation	d.f.	Σy^2	Σx^2	Σxy	Adjusted d.f.	 S.S.	 M.S.	 F.	 Hyp.
Within subclasses	100	9,927.9616	9,106.6924	6,889.0	99	4,716.5941	47.64		
Between levels	1	106.0096	71.1154	86.83	1	15.3405	15.34	--	Accept
Between methods	1	1,040.7788	26.0	-164.50	1	1,304.5395	1304.54	27.38	Reject

$$\frac{\bar{Y}_1{}' - \bar{Y}_2{}'}{\sqrt{MS_W{}' \left[\frac{1}{n_1} + \frac{1}{n_2} + \frac{(\bar{X}_1 - \bar{X}_2)^2}{SSw(x)}\right]}}$$

with degrees of freedom km - k - 1 where k equals the number of treatments or levels, and m is the number of observations in each cell.

Since the difference between any two of the four subclasses represents just one of six possible comparisons, the probability level required for significance is not set at 1 in 100 but 1 X 6 in 100, or .0017. A situation of this type is described by Johnson[15]. These tests were run on each of the six possible comparisons. The experimental group proved to be significantly different from the control group at the fifth grade level, but not at the fourth grade.

Matching of audio and visual stimuli - The test for interaction on data from Test 2 yielded an F of 5.0 so that the null hypothesis of no interaction between treatment and level remained in doubt (.05 > P > .01). However, the variation due to the interaction effect was not pooled with the residual error to test the hypothesis of no difference in means.

The analysis further revealed there was no significant difference between the means of the experimental and the control treatments with respect to the skills measured by this test (F = 2.83). Due to the fact that it was not reasonable to assume there was no interaction between the main effects of treatment and level, it was advisable to proceed in the same manner as before to test between individual subgroups.

No significant differences were discovered when the subclasses were compared with each other. It should be noted that the value of "t" was found to be 2.76 when the experimental and control groups at the fifth grade level were tested. The hypothesis of no difference between their adjusted means must, therefore, remain in doubt (.01 > P > .0017).

Music preference - The analysis and covariance showed that it was possible to assume that the main effects of treatment and level with respect to the scores on the Keston Test were independent of each other. Since the variation due to interaction was not significant, it was possible to pool it with the variation already identified as residual.

The analysis further showed that the test discriminated significantly between the experimental and control classes. F was computed to be 13.62 (P < .01). Thus, it was concluded that, within the limitations of the experiment, the combined approach of vocal and instrumental music training was superior to vocal alone when music preference as measured by this test was used as a criterion. The analysis showed this to be true at the fourth and fifth grade levels so it was not necessary to proceed beyond this point as was true in the other two instances.

Other Data

In order to reveal trends in the attitudes of the children's parents as to the value of combined vocal and instrumental training, a questionnaire canvass was made. Each set of parents of the children in both experimental control subclasses was requested to answer a twelve-item questionnaire when the experiment was nearly completed. A one hundred percent response was obtained since the questionnaires were taken home and returned by the children.

The response totals for each item were tabulated and then tested by the chi-square technique and significance accepted at the five percent level. Not all the items discriminated between the "experimental parents" and the "control parents" and there was some variation between the two levels. However, it was apparent that the parents of both fourth and fifth grade pupils accepted the program.

More of the parents of the children in the experimental groups felt their children had been stimulated to continue the study of their music than those of the controls. Another significant finding was that the experimental parents felt that the introduction of instrumental music to the curriculum would not tend to place too much emphasis on music in relation to the rest of the school program. The control parents' responses were significantly different at the five percent level on both of these points. Thus, in general, it appeared safe to assume that the parents of the children who had participated in the experimental course of study favored it.

Toward the end of the school term, the children in each of the four subclasses were given an alphabetized list of all the subjects included in their curriculum. The classroom teacher in every case administered the questionnaire and the children had no indication that the main object of the device was to discover the ranking given to music. The instruction given to the pupils by the classroom teacher was that the children rank their subjects in the order in which they liked them.

To be sure, the information thus received was not an absolute index of their preference for music. It was, rather, a relative measure because the individuals were forced to compare music with the other subjects in the curriculum. However, this method seemed preferable to the more commonly-used rating scale device, particularly at these grade levels. The chi-square test was used for determining the significance of the difference in the distribution of ranks assigned. A significant difference in favor of the experimental class was found at the fifth grade level (P < .001) but not at the fourth.

Implications

The fifth grade experimental class attained significantly higher scores on Test 1 and the Keston Test than the fifth grade control group and the null hypothesis remained in doubt on Test 2. The pupils in this class also ranked music significantly higher on the subject preference list than the other fifth grade. The fourth grade experimental subjects did not do as well, although Table I shows the same trend in both the fourth and fifth grades.

These findings would seem to suggest that it might not be advisable to introduce a combined vocal-instrumental curriculum until the fifth grade. At least, the criterion of achievement in reading and understanding music notation indicates such a conclusion. On the other hand, analysis of music preference scores showed that the experimental method produced effective results in both fourth and fifth grade experimental classes. Thus it is possible that fourth grade children are learning skills and attitudes not measured in this experiment. Further study and the use of other criteria would help supervisors to make such decisions if a program of this type were to be adopted.

The analysis of the data collected strongly indicates that the integration of both vocal and instrumental instruction would tend to develop within the pupils a broader base of musical comprehension than can be obtained through vocal training alone. Responses to questionnaire devices lead one to believe that such a program is more interesting for children and that parents would be satisfied with the arrangement.

However, it would be of great interest to know whether or not such a course of study would have lasting effects. For instance, would it be advisable to extend this instruction over a longer period of time? It may be that the desired outcomes of a music education are of such a complex nature that a one-year combined vocal-instrumental program might not be sufficient. Again, the effects of this dual instruction may not be revealed until later years so that the findings of this study do not give a true picture of the changes in the subjects.

To answer some of these questions, the writer has collected data on all of the measures on the same pupils used in this study one school year after the experiment ended. Analysis of these data will be of value because the results will be more widely applicable for the music educator. The findings of this follow-up study will be published in the near future.

FOOTNOTES

1. Arthur S. Otis. Otis Quick-Scoring Mental Tests, Alpha Test, (New York: World Book Company, 1936).
2. Florence Goodenough and John E. Anderson. Experimental Child Study, (New York: Century Company, 1931), pp. 501-512.
3. Cyril J. Hoyt and Clayton L. Stunkard. "Estimation of Test Reliability for Unrestricted Item Scoring Methods", Educational and Psychological Measurement, XII, (Winter, 1952), pp. 756-58.
4. Morton J. Keston. An Experimental Evaluation of the Efficacy of Two Methods of Teaching Music Appreciation, Unpublished Ph. D. Thesis, University of Minnesota, 1949.
5. Palmer O. Johnson. Statistical Methods in Research, (New York: Prentice-Hall, Inc., 1949), pp. 176-177.
6. H. F. Spitzer and others. Iowa Every-Pupil Tests of Basic Skills, Elementary Battery, Form L (Boston: Houghton Mifflin Company, 1945).
7. John W. Beattie and others. The American Singer, (New York: American Book Co., 1946).
8. Mabelle Glenn and others. The World of Music, (Boston: Ginn and Company, 1936).
9. Joseph E. Skornicka and Bergeim, Joseph. The Boosey and Hawkes Band Method, (New York: Boosey and Hawkes, 1947).
10. Palmer O. Johnson. op. cit. pp. 330-33.
11. Palmer O. Johnson. op. cit. pp. 82-86.
12. Ibid. pp. 160-62.
13. E. F. Lindquist, Design and Analysis of Experiments in Psychology and Education, (Boston: Houghton Mifflin Company, 1953), p. 327.
14. Alexander M. Mood, Introduction to the Theory of Statistics, (New York: McGraw Hill Book Company, Inc., 1950), p. 356.
15. Palmer O. Johnson, op. cit. p. 234.

A COMPARISON OF VERBAL AND PICTORIAL SELF-RATING-SCALE CATEGORIES

JULIAN C. STANLEY*
University of Wisconsin

DO VERBAL reference points on a self-rating scale result in different marking from pictorial ones?

Chester E. Evans[1]** has devised two forms of a "personal satisfaction" inventory which are virtually identical except that the five categories of response on one are words (very satisfied, satisfied, neutral, dissatisfied, very dissatisfied), while on the other they are five smiling, neutral, or scowling facial expressions of a man. Sample items from the inventory are: the city in which you live, your last boss, our foreign policy.[2]

Design of the Study

In order to ascertain whether or not the two sets of reference point anchors have different effects, the writer administered inventories twice to each of 48 college students of both sexes in two required courses, as outlined in Table I. In the "child psychology" (CP) class, 24 individuals were assigned at random to alternate seats in four rows. All six persons in the first row filled out the "words" inventory, followed after a 15-minute period of irrelevant activity by the "faces" inventory. Those in the third row (even-numbered rows were left vacant) received "faces" and then "words." The fifth row had "words" followed by "words," and the seventh "faces" followed by "faces." Procedure was similar in the "educational psychology" (EP) class, except that because of shorter rows, assignments were made to alternate seats in eight rows. Rows 1 and 5 received inventories in words-faces order, 2 and 6 faces-words, 3 and 7 words-words, and 4 and 8 faces-faces. More persons than are shown in Table I were tested, but enough were eliminated randomly to leave six in each class-design-order group for the statistical analysis.

Few students realized that any order other than the one they experienced was being used. The second inventories were not mentioned until 15 minutes after the first ones were turned in, during which interval regular class activities went on. The EP class ended at 11:50 a. m. and the other began at 1:20 p. m. the same day; apparently, no information concerning the experiment leaked out during this 90 minutes.

Satisfaction inventory scores, the dependent variable, consisted of mean item scores expressed on a 0 (very dissatisfied or heavy scowl) to 4 (very satisfied or broad smile) scale, multiplied by 100 to eliminate decimals. Since a few of the 50 items did not apply to many of these students (e. g., "your husband or wife"), the divisor was usually slightly less than 50. The 96 three-digit scores ranged from 178 (100 = dissatisfied, 200 = neutral) to 328 (300 = satisfied), with a grand mean of 247.

Four Ways to Analyze the Scores

The Table I scores may be analyzed in at least 10 partially different ways, but we need be concerned with only four of these.[3] Our chief interest is in the "anchor" ("words" vs. "faces") main effect. This can be tested in three different ways: by considering only scores from the first inventory; as in Table II; by treating the two non-crossover designs (1) together, as in Table III, where "anchor" is a between-individuals main effect; and by treating the two crossover designs (2, 3, 4, 5, 6, 7) together, as in Table IV, where "anchor" is a within-individuals main effect.

The first of these, in Table II, is the most powerful, having 44 d. f. for the "error" mean square, and it also provides a statistical test for the "class" main effect and the "class X anchor" interaction.

The pooled non-crossover design analyzed in Table III tests the same three sources of variation as the Table II procedure but with just 20 d. f. for "error"; in addition, it tests the "sequence" main effect and the three interactions of "sequence" with "class" and "anchor."

The Table IV replicated latin-square design involves administering both inventories to each student, unlike the other two methods, so that scores for "words" are correlated with those for "faces."

Thus with respect to "anchor" differences, Tables II, III, and IV answer three questions:

1. If two random groups are tested, one with

---

* The writer is indebted to Professor Palmer O. Johnson for several helpful suggestions concerning the reporting of this study.

**References and footnotes will be found at end of this article.

TABLE I

OVERALL DESIGN FOR COMPARING "WORDS" AND "FACES" ANCHORS

Class	Design	Order	Individual	First Inventory	Second Inventory
"Child Psychology" (CP)	Crossover	Words, Faces (Words first)	1 . 6	Words . Words	Faces . Faces
		Faces, Words (Faces first)	7 . 12	Faces . Faces	Words . Words
	Non-Crossover	Words, Words (Words first)	13 . 18	Words . Words	Words . Words
		Faces, Faces (Faces first)	19 . 24	Faces . Faces	Faces . Faces
"Educational Psychology" (EP)	Crossover	Words, Faces (Words first)	25 . 30	Words . Words	Faces . Faces
		Faces, Words (Faces first)	31 . 36	Faces . Faces	Words . Words
	Non-Crossover	Words, Words (Words first)	37 . 42	Words . Words	Words . Words
		Faces, Faces (Faces first)	43 . 48	Faces . Faces	Faces . Faces

TABLE II

ANALYSIS OF VARIANCE FOR THE 48 INVENTORIES ADMINISTERED FIRST

Source of Variation	Sum of Squares	d. f.	Mean Square	F
Between classes	4, 408. 33	1	4, 408. 33	5. 16*
Between anchors	33. 33	1	33. 33	
Class × anchor	456. 33	1	456. 33	
Within class-anchor groups**	37, 597. 00	44	854. 48	
Total	42, 495. 00	47		

* Significant beyond the . 05 level.

**Bartlett's test yields a corrected chi square of 12. 71 with 3 d. f., which is significant beyond the . 01 level, for the four class-anchor-group variance estimates whose sums of squares are pooled to yield the error term. The estimates are 486, 305, 512, and 2114.

TABLE III

ANALYSIS OF VARIANCE OF THE TWO NON-CROSSOVER DESIGNS TOGETHER

Source of Variation	d. f.	Mean Square	F*
Class	1	6, 302. 08	3. 33
Anchor	1	2, 214. 08	1. 17
Class × anchor	1	1, 776. 33	
Individuals within class-anchor groups	20	1, 894. 98	118. 50**
Sequence	1	2. 08	
Class × anchor × sequence	1	36. 75	2. 30
Class × sequence	1	0	
Anchor × sequence	1	0. 33	
Individual × sequence within class-anchor groups	20	15. 99	
Total	47		

* Computed from mean squares carried to four decimal places.

**Significant beyond the . 001 level.

TABLE IV

ANALYSIS OF VARIANCE OF THE TWO CROSSOVER DESIGNS TOGETHER

Source of Variation	d.f.	Mean Square	F
Class	1	3,605.33	2.25
Order	1	2,106.75	1.31
Class × order	1	280.33	
Individuals within class-order groups	20	1,603.68	33.65*
Sequence	1	8.33	
Order × sequence (anchor)	1	96.33	2.02
Class × sequence	1	44.08	
Class × anchor	1	102.08	2.14
Individual × sequence within class-order groups	20	47.66	
Total	47		

*Significant beyond the .001 level.

TABLE V

OVERALL ANALYSIS OF VARIANCE OF SCORES IN TABLE I DESIGN

Source of Variation	d.f.	Mean Square	F
Between individuals			
Class	1	9,720.38	5.56*
Design	1	805.04	
Order (W first vs. F first)	1	0.67	
Class × design × order	1	322.67	
Class × design	1	187.04	
Class × order	1	1,734.00	
Design × order	1	4,320.17	2.47
Individuals within CDO groups	40	1,749.32	54.97**
"Within" individuals			
Sequence	1	1.04	
Class × design × order × sequence	1	8.17	
Class × design × sequence	1	22.04	
Class × order × sequence	1	130.67	4.11***
Design × order × sequence	1	42.67	1.34
Class × sequence	1	22.04	
Design × sequence	1	9.38	
Order × sequence	1	54.00	1.70
Individual × sequence within CDO groups	40	31.82	
Total	95		

* Significant beyond the .025 level.
** Significant beyond the .001 level.
***Significant at the .05 level.

TABLE VI

VARIANCE ESTIMATES THAT WERE POOLED TO YIELD THE TWO ERROR TERMS IN TABLE V

Class-Design-Order Group	d. f.	Mean Square	
		Between Individuals	i × s
CP, Crossover, WF	5	981.53	50.93
CP, Crossover, FW	5	329.35	107.88
CP, Non-C, WW	5	1,074.28	18.95
CP, Non-C, FF	5	576.80	4.93
EP, Crossover, WF	5	1,308.33	21.93
EP, Crossover, FW	5	3,975.48	9.88
EP, Non-C, WW	5	624.13	12.60
EP, Non-C, FF	5	5,304.68	27.48
Corrected chi square (Bartlett)		15.27	15.38
Chi square at .05 level (7 d.f.)		14.07	14.07

"words" and the other with "faces," will their mean scores differ significantly?

2. If two random groups are tested twice each, one with "words" both times and the other with "faces," will the means of their two summed scores differ significantly?

3. If two random groups are tested, one with "words" followed by "faces" and the other with "faces" followed by "words," will the mean for "words" differ significantly from that for "faces?"

Usually we are not trying to answer the second question. The first and third questions are essentially equivalent, so it would seem most suitable with a fixed number of subjects to employ only the crossover design for all members in each class, thus reducing the "error" term for the anchor main effect by taking into account the high correlation between "words" and "faces," besides providing as many degrees of freedom for "error" as the Table II analysis. The crossover design requires retesting each subject and a considerable degree of control, but it yields correlational information not available when each person is tested only once. Also, it will often be more efficient, both statistically and financially, to secure 96 scores from 48 examinees than 96 scores from 96 examinees.

However, the Table I mixture of crossover and non-crossover designs permits us to test differences among the eight *r*'s and thereby to see whether the "unalike anchor" reliability coefficients differ significantly from the "test-retest" coefficients. They should not if the two forms are to be regarded as interchangeable.

The Table I design cannot be analyzed as a whole without losing the "anchor" classification by dealing with "order" ("words inventory taken first" vs. "faces inventory taken first"), as shown in Table V.

Results

"Words" and "faces" anchors do not differ significantly from each other in any of the analyses (Tables II, III, and IV), nor does "anchor" interact significantly with class or sequence. In Tables II and V the classes differ significantly beyond the .05 level, and in Table V the interaction of class, order, and sequence appears to be significant. However, because of the large number of significance tests run (15 main ones in Table V) and the heterogeneous "error" variance shown in Table VI, this apparently significant C × O × S mean square may be attributable to chance; if real, its meaning is not clear.

The eight Pearsonian *r*'s, each based upon six pairs of scores, range from .55 for the CP-crossover-faces first group to .9961 for the EP-crossover-faces first group. The Fisher-*z* equivalents of these *r*'s have an estimated population variance of 0.6182, compared with the theoretical variance of 0.3333 for random-sample *z*'s based upon six pairs. The ratio of 0.6182 to 0.3333 is a non-significant *F* of 1.85 with 7 and infinite degrees of freedom.[4] Thus the eight *r*'s do not differ significantly from each other.

Discussion

The two chief findings from the analyses outlined above are high correlation between the inventories and somewhat greater personal satisfaction in the EP class than in CP. Perhaps the CP class members, chiefly first-semester juniors enrolled for their initial course in the School of Education, were under more tension than the first-semester seniors in the EP class, who had already been self-screened for malcontents. Several alternative hypotheses are equally plausible, of course.

"Words" and "faces" means do not differ significantly for the college students employed as subjects in this study.[5] Scores on the two inventories have similar variability and appear to intercorrelate highly. Obviously, it is possible that other verbal categories would produce reactions different from the faces, since the word "very" may be quite indeterminate. It should be interesting to compare various rating-scale markers, such as the numbers +2 through -2 or pips on a line, with the word and face anchors.

Although the satisfaction inventories originated in industrial situations, results of this study should not be over-generalized. There may be marked differences between the reactions of advanced college students and industrial workers. Some of the designs discussed here should be suitable for experimentation with various non-college groups.

Summary

Using with 48 college students an experimental design involving both crossover and non-crossover principles, the writer shows by four analyses of variance that two "personal satisfaction" inventories devised by Chester E. Evans which have words as rating-scale reference points in one and faces in the other yield equivalent mean scores, differentiate between two college classes, and intercorrelate approximately .96.

REFERENCES

1. Block, J., and others. "Testing for the Existence of Psychometric Patterns," *Journal of Abnormal and Social Psychology*, XLVI (1951), pp. 356-359.
2. Edwards, A. L. "Homogeneity of Variance and the Latin Square Design," *Psychological Bulletin*, XLVII (1950), pp. 118-129.

3. Edwards, A. L. Experimental Design in Psychological Research (New York: Rinehart and Co., 1950).
4. Grant, D. A. "The Latin Square Principle in the Design and Analysis of Psychological Experiments," Psychological Bulletin, XLV (1948), pp. 427-442.
5. Grant, D. A. "The Statistical Analysis of a Frequent Experimental Design," American Journal of Psychology, LXII (1949), pp. 119-122.
6. Stanley, J. C. "Gellerman's Complex Crossover Design," Journal of Consulting Psychology, XVIII (1954), p. 380.
7. Stanley, J. C. "Statistical Analysis of Scores from Counterbalanced Tests." See article in this issue of Journal of Experimental Education, pp. 187-207.
8. Weitz, J. "A Neglected Concept in the Study of Job Satisfaction," Personnel Psychology, V (1952), pp. 201-205.

FOOTNOTES

[1] Employe Research Section, General Motors Corporation, Detroit 2, Michigan.

[2] Twenty-seven of the 50 items were originated by Joseph Weitz (8). Two other sources of information concerning the history and development of the satisfaction inventories are Technical Bulletins No. 10 and 16, Employe Research Section, General Motors Corporation: Theodore Kunin, "A Study in the Placement of Faces Along Subjective Continua," June, 1950, and "The Weitz-Evans Community Checklist," January, 1954. Kunin's report, an M. A. thesis at Western Reserve University, is scheduled for condensed publication in a forthcoming issue of Personnel Psychology. The present writer is indebted to Dr. Evans for furnishing the inventories and calling his attention to their antecedents.

[3] The composite design of Table I involves two crossover and two non-crossover designs, each of which four might be subjected to an analysis of variance separately, and two classes, which might also be analyzed apart from each other. The writer has done these six analyses for his own information and will be glad to correspond with interested persons concerning them, but they are not discussed in this article.

[4] This procedure is equivalent to the chi-square test outlined by Edwards (3:135).

[5] Since no population was defined and random samples taken from it, conclusions must be limited to these 48 subjects or to some hypothetical population for which they might be conceived to constitute a random sample.

THE INTERPRETATION OF THE STANDARD ERROR OF MEASUREMENT

C. H. PATTERSON
Veterans Administration Regional Office
St. Paul, Minnesota

THE USE of the standard error of measurement as a measure of the reliability of test scores has been recommended by some test experts (18:222). Since the standard error of measurement applies to a single score, and is stated in terms of test score units, it is much more useful in counseling individuals than the reliability coefficient. Although Kelley pointed out in 1921 (13) that the standard error of measurement is relatively independent of the range of scores (or ability), a factor which has considerable effect on coefficients of reliability, this apparently is true only under certain conditions (17, 20, 21).

Recently the writer, in discussing the reliability of the Wechsler-Bellevue Intelligence Scales (23), had occasion to refer to the standard error of measurement of an IQ obtained with this instrument. Reference was made to several standard statistical texts. It was discovered that many texts make no reference to the standard error of measurement (3, 25, 26, 28, 30), while others give no definition or interpretation of it. It is interesting that Cronbach's recent book on psychological testing (1) makes no mention of this measure of reliability. In the recent volume on educational measurement, edited by E. F. Lindquist, R. L. Thorndike, in his excellent chapter on reliability (27), refers to this statistic, but does not discuss its meaning or use.

In view of the lack of information and the apparent disagreement of interpretation, the writer felt that other test users and counselors would be interested in a discussion of the meaning and interpretation of this measure of reliability. The controversy between Guilford and Garrett (5, 6, 9) apparently has not resolved the issues. Although it appeared that Garrett agreed fundamentally with Guilford, he persisted in maintaining his disagreement, and in asserting the equivalence of two different interpretations.

The standard error of measurement, variously designated as σ_e, $\sigma_{1\infty}$, or $\sigma_{1.\infty}$, σ_{1w}, or σ_m is given as

$$\sigma_{1\infty} = s_1\sqrt{1 - r_{11}} \tag{1}$$

where s_1 is the standard deviation of the distribution of test scores, and r_{11} is the reliability coefficient of the test, preferably based on the correlation of alternate (equivalent) forms. This formula gives the standard error of an actual observed score; the standard error of measurement is the standard deviation of the distribution of the scores an individual would make on a large number of testings with equivalent forms. The mean of this distribution of obtained scores would be the individual's "true score" on the test.

A number of the various interpretations of this measure will be quoted and discussed.

1. One type of interpretation consists of statements concerning the true score of individuals obtaining a specified observed score. These statements are in the form of the probability of the true score falling within specified limits of the obtained score. It is stated, usually, that the true score of an individual making a specified score is almost certain to lie within three standard errors of the obtained score, or, that the chances are approximately 68 out of 100 that the true score lies within ± one standard error of the obtained score. Peatman, for example, referring to the confidence limits of an IQ of 109, given a standard error of 4.5, states: "... thus it is likely that persons with an IQ score of 109 have parameter (i.e., true) scores whose values lie between 100.0 and 118.0; and we can be quite confident that the parameter values will not be less than 95.5 or greater than 122.5" (24:379). While not an exact probability statement, this appears to be of the form stated above.

2. Guilford (8:414) points out that such statements imply the prediction of the true score from the observed score. This would require the standard error of estimate of a true score. In another place (10:480) he states: "Any obtained score does not tell us what the corresponding true score is, but with knowledge of $\sigma_{1\infty}$ we can have a degree of confidence that the true score cannot be very far away." This implies that no exact statement concerning the true score can be made from an observed score. As Lindquist points out (18:221 footnote), the standard error of measurement is the standard error of estimating an observed score on a test from the corresponding true score. Thus the proper interpretation of the standard error of measurement is in terms of the distribution of obtained scores for a given true score. Guilford (8:414) gives as the correct type of state-

ment: "If a certain individual has a true score of 50 points in the test, then we may expect two-thirds of all his actual scores to lie between 46 and 54"(given a $\sigma_{1\infty}$ of 4). Lindquist agrees with this interpretation (18:221-222).

Variations of this type of interpretation are in terms of the probabilities of any observed score falling within certain limits of the true score, or the probabilities of the obtained score differing from the true score by a given amount, say, one or more standard errors, or probable errors. For example, Kelley (15:79) writes: ". . . . the chances are fifty in one hundred that the pupil's obtained score differs from his true ability score by an amount greater than [one PE]." Garrett also follows this latter interpretation. He says (using a PE of 3): "This result may be interpreted to mean that the chances are even (50 in 100) that the obtained score of any individual in the group . . . does not differ from his true score by more than ± 3 points" (7:321).

It will be noted that these interpretations differ from those of Guilford and Lindquist in that they make statements concerning an individual score. That is, they assign a probability to an individual score, rather than to a distribution of scores. The error of this procedure will be discussed later.

Gulliksen also follows the interpretation in terms of obtained scores, in the manner of Guilford and Lindquist. He states that, ". . . . in general, no probability statements can be made to apply to all persons who make a given observed score. We can only make the statement the other way around. For all persons with a given true score, the probability is over .997 that the observed score will lie within plus or minus three times the standard error of measurement from that true score" (11:20).

3. However, another interpretation appears possible. While it may not be permissible to make a probability statement regarding all persons with a given observed score, it is possible to make statements regarding a particular individual with a given observed score. There appear to be three methods of doing this. The first one to be discussed is the method used by Kelley and others, including McNemar (24) and Garrett (7:320), in which the observed score is taken as an estimate of the true score. While Kelley defines the standard error of measurement as "the standard deviation of single scores for a given fixed value of the true score," this is equivalent, for him, to "the standard deviation of errors of estimate when the single score is taken as evidence of a true score" (15:171). On this basis, Kelley makes interpretations such as the following (given a score of 70 and a standard error of 7.72): "There is one chance in six that John's true ability lies below 62.3, so that there are four chances in six that it lies between 62.3 and 77.7" (15:172). In another place Kelley writes: "$\sigma_{1\infty}$ may correctly be used to indicate the standard error when x_1 (an obtained score) is estimated from a knowledge of x_∞ (a true score)(a process actually never performed), or when x_∞ is estimated by using non-regressed x_1" (16:403).

These statements concerning the true score are similar to that of Peatman quoted before. However, Peatman does not specify that the observed score is taken as evidence of the true score. There is a difficulty in the nature of the statements made which will be referred to later. The correct form of statement regarding a true score on the basis of the standard error of an observed score which is taken as evidence of a true score is of the type that for a given observed score, the corresponding true score will fall within a given interval, determined by the observed score and a multiple of its standard error according to the confidence coefficient desired, and we shall be right in the statement that the interval will cover the true score in a specified proportion of the cases corresponding to the confidence coefficient selected. It should also be stated that the confidence interval limits are random variables.

4. There is a second way in which statements about true scores may be made. While it is true that the observed score is an estimate of the true score, it is not the true score, nor the best estimate of the true score. Using an observed score as the estimate of the true score and applying the standard error to the observed score as if it were the error of the true score (as Kelley does above) may be confusing. It may be argued that the best estimate of the true score is one obtained as a result of prediction based upon the regression of the true scores upon the observed score, rather than accepting the observed score as the estimate of the true score.

The standard error of estimate in predicting a true score from an observed score by means of a regression equation is given by

$$\sigma_{\infty 1} = s_1\sqrt{1 - r_{1I}} \qquad \sqrt{r_{1I}} \qquad (2)$$

$$\sigma_{\infty} = s_1\sqrt{r_{1I}(1 - r_{1I})}$$

where s_1 is the standard deviation of the observed scores and r_{1I} is the coefficient of reliability of the observed scores. Since the reliability coefficient is always less than one, it is clear that the standard error of estimate will always be less than the standard error of measurement. That is, the error of a regressed score taken as an estimate of true score is always less than the error of an obtained score taken as an estimate of a true score.

The usual interpretation of the error of estimate is in terms of the confidence interval including the true score. Thus it may be stated that (for additional samples) in 99 percent of the cases the true score falls within the interval ± three standard errors above and below the estimated true score.

The standard error of estimate of a regressed score is seldom used as a measure of reliability. Guilford writes: "There is nothing to be gained by [predicting true scores] for the predictions would be no more accurate than the scores from which they were obtained" (10:479). Kelley, however, states that "to secure an estimate of true ability, [this] process is in all cases the best, but if reliability is high, it is not sufficiently better.... to warrant the extra labor" (15:181). Gulliksen agrees that "no practical advantage is gained from using the regression equation to estimate true scores" (11: 45).

5. No exact probability statement can be made concerning the group of individuals with a given observed score. However, *reasonable limits* may be set for the true score of an individual with a given observed score. Gulliksen describes the procedure as follows: "We may also indicate a method for assigning a probability value to the statement: 'If a person's observed score is 65, his true score probably lies between 50 and 80 ($\pm 3\sigma_{1\infty}$).' Note that no probability value is given. We cannot say that for all persons whose observed score is 65, the probability is greater than .99 that the true score is between 50 and 80. However, consider the statement: 'This person's true score lies between 50 and 80'. If the person's true score were known, the statement could, for a given person, be classified as *true* or *false*.... For each of the persons whose observed score is known, such a statement can be made.... For all persons in the distribution, it will be found that the statement regarding limits is true over 99 percent of the time and is false less than three times out of a thousand. In other words, if all the cases were considered, a probability can be attached to the truth or falsity of the statement that 'true score is included within the specified limits.' However, if we limit the statement to persons with any *given observed score*, no assertion regarding probability can be made. ...We say then that for a person whose observed score is 65, *reasonable limits* for his true score are 50 to 80" (11:17-20).

Statements such as that of Peatman, quoted earlier, may be technically correct, since they use terms such as "likely" and "quite confident." However, they are misleading in that they imply exactness to the reader. All that can be done is to set "reasonable limits" for the true score, without assigning exact probabilities.

6. A method of assigning exact probabilities rather than "reasonable limits" to statements referring to a group of individuals with a given obtained score would appear to be possible through the use of tolerance limits.[1] Tolerance limits determine an interval which will include at least a specified proportion, P, of the distribution, with a specified confidence.

Tolerance limits are of the form

$$\overline{X} \pm Ks \qquad (3)$$

where $\overline{X}$ and s are the mean and the standard deviation of the sample, and K is one-half of the tolerance interval, which may be determined by reference to a table in which values of K are given for specified values of P, of the confidence coefficient or level desired, and of N (4:102-107; also reproduced in 2, Table 16 in the Appendix). In the present application $\overline{X}$ would be the obtained score (taken as an estimate of the mean of the distribution of obtained scores in repeated testing, i.e., of the true score), and s is the standard error of measurement (taken as an estimate of the standard deviation of the distribution of obtained scores, i.e., of an estimate of the standard error of the obtained score taken as an estimate of the true score). In repeated testing on equivalent tests, the group of individuals with a given obtained score on the first test would obtain varying scores on the other tests, and the standard error of measurement would also vary from test to test. Tolerance limits enable one to state, with a specified degree of confidence, the interval which will include a specified proportion P of the obtained scores.

As an example of the procedure, let us take an obtained score of 50, with a standard error of 4. It is desired that tolerance limits be set which will include 95 percent of the obtained scores in future samples, or on equivalent tests, with a confidence of .99. The N would be the number of cases entering into the determination of the standard error of measurement, say, 100 in this example. K would in this case be 2.355. The interval would then be $50 \pm K\sigma_{1\infty}$, or 40.58 -59.42. We can thus be sure, at the .99 confidence level, that in repeated testing of this sample of 100 cases (or in other samples of the same size), 95 percent of the obtained scores (or estimates of the true scores) of the group with an original obtained score of 50, will fall within this interval. That is, in 99 percent of the equivalent tests, 95 percent of the obtained scores will be included within the interval (or, still differently, 99 percent of the tolerance limits determined from repeated tests will include 95 percent of the scores). If we desired to include 99 percent of the scores in 95 percent of the tests, the limits would be 38.26 - 51.74. By increasing P we can, with a corresponding increase in the size of the interval, include a greater proportion of the scores.

---

[1] The possibility of applying tolerance limits in this situation was suggested by Dr. Palmer O. Johnson, University of Minnesota.

Finally, it is necessary to consider the nature of probability statements regarding confidence intervals within which a given parameter such as a true score, lies. The true score is a fixed value, not a random variable. (In Fisher's concept of fiducial limits, the parameter is the same as in the case of confidence intervals.) Since it is fixed, it either is or is not within the confidence interval in a particular case, and the probability is either unity or zero, respectively. The probability therefore applies to the statement, which will be true in the proportion of cases specified by the confidence coefficient, and not to the individual score[2] (12:111; 22:221-222; 29:258-261).

It would appear to follow that any statement concerning an individual score must take the form given above. Thus, the correct statement concerning a given observed score would be of the type: "This individual's observed score lies within plus and minus one standard error of his true score, and we shall be correct in this statement in 68 cases out of 100," or (taking the obtained score as an estimate of the true score), "This individual's score lies within plus or minus one standard error of his observed score, and we shall be correct in this statement in 68 out of 100 cases."

Accepting this frequency theory of probability, it appears possible, however, to make a probability statement concerning a group of individuals, or a group of scores, when the nature of the statement is appropriate. (1) As an example, "for the group of persons whose true score is 50, over 99 percent of the observed scores will be between three standard errors above and below the true score" (11:17). (2) In the case of true scores estimated by means of the regression formula, the correct statement would be that for all cases with the same estimated true score, the actual true score falls within three standard errors of the estimated true score over 99 percent of the time. (3) For a group of individuals with a given obtained score, in repeated testing the obtained scores will almost certainly fall within three standard errors ("reasonable limits") of the corresponding true score. (4) For a group of individuals with a given observed score, in repeated testing a specified proportion of the obtained scores will fall within given tolerance limits with a specified probability, i. e., in a specified proportion of the samples of equivalent tests.

Summary

A survey of a number of standard texts in statistics indicated some disagreement in the interpretation of the standard error of measurement. Representative interpretations were quoted and classified under several types. Two main types of difficulty appear in interpreting the standard error of measurement. The first concerns the problem of making statements concerning true scores on the basis of the standard error of observed scores. Guilford and Lindquist prefer to make interpretations only in terms of the distribution of obtained scores in repeated testing, or of the distribution of the observed scores of individuals with a given (assumed) true score. On the other hand, various writers make statements concerning the true score on the basis of an observed score and its standard error. They do this by implicitly or explicitly taking the observed score as the estimate of the true score, and using the standard error of the observed score as being equivalent to the standard error of a true score. While the observed score is an estimate of the true score, it is not the true score, nor is it necessarily the best estimate of the true score. It would appear that this type of interpretation is therefore questionable. Statements concerning the true scores of individuals with a given observed score should be based on the regressed score and its standard error of estimate. While the effort involved in this procedure of estimating true scores may not be justified, statements concerning the true scores associated with observed scores should be limited to such regressed scores. Rigorous logic would indicate, therefore, that statements should be of the type given by Guilford and Lindquist. Nevertheless, use of the observed score as the estimate of the true score is commonly accepted, and appropriate statements concerning the true scores of individuals with a given observed score taken as evidence of the true score are given.

A type of interpretation which makes statements about the group of individuals with a given observed score was mentioned. No simple exact statement can be made about all the individuals with a given observed score. A type of exact probability statement based on the application of tolerance limits has been suggested.

---

[2] Another way of stating it would be to say that the confidence interval computed from repeated sample estimates will cover the true score in the proportion of cases specified by the confidence coefficient.

This problem is of little practical interest, however, since we are usually not interested in making statements about groups of individuals with a given observed score.

The second main difficulty concerns the type of statement which can be made. Probability applies to the statement, which is true in a given proportion of cases, and not to the score or parameter, which is fixed, i. e., it either is or is not within the stated limits in a particular case. It cannot be stated, therefore, that the probability is two-thirds that a given individual's true score lies within plus or minus one standard error of his observed score taken as an estimate of his true score. For all the individuals in the distribution, the probability is two-thirds that their true scores lie within plus or minus one standard error of their observed scores, i. e., it can be stated that "this individual's true score lies within plus or minus one standard error of his observed score", and we shall be correct in two-thirds of the cases in the distribution.

This type of statement led to a consideration of the nature of probability statements which are appropriate in other interpretations. It is suggested that, following the frequency theory of probability, statements concerning an individual score must be of this type, the probability applying to the truth or falsity of the statement, not to the score. For groups of individuals, or for the distribution of scores of an individual, probability statements can be made in the usual fashion, and the types of appropriate statements are suggested. The use of tolerance limits makes it possible to make statements concerning the interval which will include a specified proportion of the scores on repeated tests of a group of individuals with a given observed score, to which an exact probability can be assigned.

REFERENCES

1. Cronbach, L. J. Essentials of Psychological Testing (New York: Harper and Brothers, 1949).
2. Dixon, W. J. and Massey, F. J., Jr. Introduction to Statistical Analysis (New York: McGraw-Hill Book Co., 1951).
3. Edwards, A. L. Statistical Analysis for Students in Psychology and Education (New York: Rinehart and Co., 1946).
4. Eisenhart, C. and others (Eds.) Techniques of Statistical Analysis (New York: McGraw Hill Book Co., 1947).
5. Garrett, H. E. "On the Interpretation of the Standard Error of Measurement," American Journal of Psychology, XLIX (1937), pp. 679-680.
6. ______ "A Rejoinder," American Journal of Psychology, XLIX (1937), pp. 683-685.
7. ______ Statistics in Psychology and Education (New York: Longmans, Green and Co., 1947).
8. Guilford, J. P. Psychometric Methods (New York: McGraw-Hill Book Co., 1936).
9. ______ "More Concerning the Interpretation of the Standard Error of Measurement," American Journal of Psychology, XLIX (1937), pp. 680-683.
10. ______ Fundamental Statistics in Psychology and Education, Second Edition (New York: McGraw-Hill Book Co., 1950).
11. Gulliksen, H. Theory of Mental Tests (New York: John Wiley and Sons, 1950).
12. Johnson, P. O. Statistical Methods in Research (New York: Prentice-Hall, 1949).
13. Kelley, T. L. "The Reliability of Test Scores," Journal of Educational Research, III (1921), pp. 370-379.
14. ______ Statistical Methods (New York: Macmillan Co., 1923).
15. ______ Interpretation of Educational Measurements (New York: World Book Co., 1927).
16. ______ Fundamentals of Statistics (Cambridge: Harvard University Press, 1947).
17. Lord, F. M. "A Theory of Test Scores," Psychometric Monographs, No. 7 (1952).
18. Lindquist, E. F. A First Course in Statistics, Rev. Ed. (Boston: Houghton Mifflin Co., 1942).
19. McNemar, Q. Psychological Statistics (New York: John Wiley and Sons, 1949).
20. ______ "Statistical Theory and Research Design," Annual Review of Psychology, III (1952), pp. 409-418.
21. Mollenkopf, W. G. "Variation of the Standard Error of Measurement," Psychometrika, XIV (1949), pp. 189-229.
22. Mood, A. M. Introduction to the Theory of Statistics (New York: McGraw-Hill Book Book Co., 1950).
23. Patterson, C. H. The Wechsler-Bellevue Scales: A Guide for Counselors (Springfield, Ill.: Charles C. Thomas, 1953).
24. Peatman, J. G. Descriptive and Sampling Statistics (New York: Harper and Brothers, 1947).
25. Peters, C. C. and Van Voorhis, W. R. Statistical Procedures and Their Mathematical Bases (New York: McGraw-Hill Book Co., 1940).
26. Rider, P. R. An Introduction to Statistical Methods (New York: John Wiley and Sons, 1939).
27. Thorndike, R. L. "Reliability," in E. F. Lindquist (Ed.) Educational Measurement

(Washington, D. C.: American Council on Education, 1951), pp. 560-620.
28. Trelear, A. E. Elements of Statistical Reasoning (New York: John Wiley and Sons, 1939).
29. Waugh, A. E. Elements of Statistical Methods (New York: McGraw-Hill Book Co., 1952).
30. Walker, H. M. Elementary Statistical Methods (New York: Henry Holt and Co., 1943).

SIMULTANEOUS EXAMINATION AND METHOD ANALYSIS BY VARIANCE ALGEBRA

WILLIAM J. MOONAN*
U. S. Naval Personnel Research Field Activity
San Diego, California

Introduction

CONSIDERABLE attention is often given to the construction of examinations which are to be utilized in some respect for educational and psychological research purposes. Frequently, these tests are used in method experiments of varying complexity. It is a matter of record that very little attention is given to the examination after or during the execution of the experiments. Frequently consistency coefficients are quoted from data obtained from extra-experimental sources, but it is more important to report these when they result from the imposition of the methods upon the experimental subjects. This is especially true, if the methods impose conditions which are somewhat different from those of the original examination standardization.

Another important matter concerns whether or not any or a set of items "interact," in the non-additive sense, with the different methods. In order to point out the relevancy to methods experiments, consider this hypothetical educational example. Suppose a 50 item test is given to two classes of equal size of arithmetic students who were taught by two different methods. If the students of the first class answered the first 25 items correctly and missed the last 25 items, while the students of the second class reversed this type of response, then the two sample means which are calculated from the total test scores of the two groups would be equal and the hypothesis of equality of the two methods would be sustained.

Certainly we would not like to conclude that the methods of teaching were equivalent since we lack the necessary basis for judging the equivalence with a test whose items so obviously interact with the methods. This may happen on a smaller scale in real educational experiments and consequently should be guarded against. We need to use an examination with items which are effected by, but do not interact with, the factors of the experiment. Tests of hypotheses for item-method interactions and consistency coefficients would be of great value in the analysis of experiments utilizing examination data.

It is the purpose of this paper to show how item-factor interactions may be tested in a 1-factor experiment where the variables are obtained as a result of an examination score and to evaluate the internal consistency coefficient and the index of internal consistency of the examination. The factor will be referred to as "methods" in the discussion which follows. These ideas may be easily generalized to other designs typically used in educational and psychological research.

The Analysis of Examination Data in a Methods Design

It is immediately recognizable that the item responses on the examination are not stochastically independent for a particular subject who participates in either method. We shall make the assumption that this dependency is measured by $\rho(I)$ which is Fisher's intra-class correlation coefficient. Thus we may write the expected value of the

$$I \sum_{m=1}^{M} S(m) = IS \quad \text{normal variables as}$$

$$(1)\ \mathcal{E}\, y(i,m,s) = \Phi(i,m) = \xi + \iota(i) + \mu(m) + \theta(i,m)$$

where $i = 1, \ldots, I$; $m = 1, \ldots, M$; $s = 1, \ldots, S(m)$.

In this equation, ξ represents the general effect, $\iota(i)$, the effect of the ith item; $\mu(m)$, the effect of the mth method; and $\theta(i,m)$, the effect of the (i, m)th interaction. The model may be reparameterized by the following linear restrictions:

$$\sum_{i=1}^{I} \iota(i)=0; \quad \sum_{m=1}^{M} \mu(m)=0; \quad \sum_{i=1}^{I} \theta(i,m)=\sum_{m=1}^{M} \theta(i,m)=0.$$

The variance of any observation, $y(i,m,s)$, is $V[y(i,m,s)] = \sigma^2$. The following table indicates this fact as well as provides the covariances, C, for various combinations of i, m, and s. The primes denote different values of the indices, except where equalities are indicated. Notice that $C[y(i,m,s),\ y(i,m,s')] = \rho(I)\sigma^2$ and that the variance, σ^2, and the covariance, $\rho(I)\sigma^2$, are assumed to be equal for all methods.

---

*The opinions expressed are solely those of the author and are in no way official; nor are they to be construed as representing those of the U. S. Naval Personnel Research Unit or Bureau of Personnel.

	i=i'		i≠i'	
	s=s'	s≠s'	s=s'	s≠s'
m=m'	σ^2	0	$\rho(I)\sigma^2$	0
m≠m'	0	0	0	0

We now follow a procedure suggested by Nandi (3). An orthogonal linear transformation is made on the y(i, m, s). The equations of transformation are:

$$(2)\quad z(h,m,s) = \left[h\,y(\overline{h+1},m,s) - \sum_{i=1}^{h} y(i,m,s)\right] \Big/ \sqrt{h\,\overline{h+1}}\,;\ h=1,\ldots,\overline{I-1}$$

$$(3)\quad z(I,m,s) = \sqrt{I}\ y(.,m,s) = \sum_{i=1}^{I} y(i,m,s)/\sqrt{I}.$$

Hence, the variables z(1, m, s), z(2, m, s), . . . , $z(\overline{I-1},m,s)$ are normally and stochastically independent. The variance of the variables is $\sigma^2[1-\rho(I)]$ since

$$(4)\quad V[z(h,m,s)] = \left\{h^2 V[y(\overline{h+1},m,s)] + V\left[\sum_{i=1}^{h} y(i,m,s)\right] - 2hC\left[y(\overline{h+1},m,s),\ \sum_{i=1}^{h} y(i,m,s)\right]\right\} \Big/ h\,\overline{h+1} = \left\{[h^2\sigma^2] + [h\sigma^2 + h\,\overline{h-1}\,\rho(I)\sigma^2] - 2h[h\,\rho(I)\sigma^2]\right\} \Big/ h\,\overline{h+1} = \sigma^2[1-\rho(I)],\ \text{and}$$

$$C[z(h,m,s),\ z(h',m,s)] = 0.$$

Also it is easy to show that $V[z(I,m,s)] = \sigma^2[1 = \overline{I-1}\ \rho(I)]$ and $C[z(I,m,s),\ z(h,m,s)] = 0$. Clearly, $\mathcal{E}z(I,m,s) = \sqrt{I}[\xi + \mu(m)]$ and is functionally independent of $\iota(i)$ and $\theta(i,m)$. Because of the orthogonal transformation,

$$(5)\quad \sum_{m=1}^{M}\sum_{s=1}^{S(m)}\sum_{i=1}^{I}[z(i,m,s) - \mathcal{E}z(i,m,s)]^2 = \sum_{m=1}^{M}\sum_{s=1}^{S(m)}\sum_{i=1}^{I}[y(i,m,s) - \mathcal{E}y(i,m,s)]^2.$$

We now break up the left side of (5) into two parts:

$$(6)\quad \sum_{m=1}^{M}\sum_{s=1}^{S(m)}\sum_{i=1}^{\overline{I-1}}[z(i,m,s) - \mathcal{E}z(i,m,s)]^2 + \sum_{m=1}^{M}\sum_{s=1}^{S(m)}[z(I,m,s) - \mathcal{E}z(I,m,s)]^2.$$

Making use of the fact that

$$(7)\quad \sum_{m=1}^{M}\sum_{s=1}^{S(m)}[z(I,m,s) - \mathcal{E}z(I,m,s)]^2 = \sum_{m=1}^{M}\sum_{s=1}^{S(m)} I[y(.,m,s) - \xi - \mu(m)]^2,$$

we are finally able to discern the following equality

$$(8)\quad \sum_{m=1}^{M}\sum_{s=1}^{S(m)}\sum_{i=1}^{\overline{I-1}}[z(i,m,s) - \mathcal{E}z(i,m,s)]^2 = \sum_{m=1}^{M}\sum_{s=1}^{S(m)}\sum_{i=1}^{I}[y(i,m,s) - y(.,m,s) - \iota(i) - \theta(i,m)]^2$$

which has $\overline{I-IS}$ degrees of freedom associated with it. The total sum of squares (5) has thus been partitioned into two parts which may be called the "within methods" sums of squares, (8), and the "between methods" sums of squares, (7).

This design has features which are similar to the "split-plot" design which is frequently used in agricultural research (see refs. 1, 4). For lack of a better name, the model being analyzed here might be termed the M-sample intra- and inter-method linear hypothesis. The hypotheses of the intra- and inter-method partitions are tested with different precisions. These precisions are estimates of $\sigma^2[1-\rho(I)]$ and $\sigma^2[1 + \overline{I-1}\,\rho(I)]$ respectively. The tests made within one partition are stochastically independent of the tests made within the other

The rank, or degrees of freedom, of the estimation space for the intra-method partition is IM-M since there exists IM-M linearly independent interaction parameters, which persist through the reduction of the reparameterized estimation space of $\mathcal{E}[y(i,m,s) - y(.,m,s)] = (i) + \theta(i,m)$ by elementary transformations. The rank for the error sums of squares for this partition is given by $\overline{I-IS} - \overline{IM-M} = \overline{I-1}\ \overline{S-M}$. The hypotheses $\iota(i) = 0$ and $\theta(i,m) = 0$ yield sums of squares which can easily be shown to have $\overline{I-1}$ and $\overline{I-1}\ \overline{M-1}$

degrees of freedom associated with them.

The rank of the estimation space for the inter-method partition is M and the rank for the error sums of squares is $\overline{S-M}$. The hypotheses $\xi = \xi(o)$ and $\mu(m) = 0$ have sums of squares with 1 and $\overline{M-1}$ degrees of freedom. As a result of these observations and some further algebra, we find the analysis of variance given by Table I.

With the aid of the statistics calculated for the entries in the sums of squares column of Table I, it is possible to test the various hypotheses listed in the first column. These tests include the desired item-method interaction test which is obtained by using the mean squares of $SS[IS]$ and $SS[S]$ as an F ratio with the degrees of freedom $(1) = \overline{I-1}\ \overline{M-1}$ and $(2) = \overline{I-1}\ \overline{S-M}$, that is, $F[\overline{I-1}\ \overline{M-1}, \overline{I-1}\ \overline{S-M}; 1 - \alpha]$.

Considerations of Examination Internal Consistency

Now we shall be concerned with a measure of the internal consistency or the intra-class correlation, $\rho(I)$, of the responses to the I items. Also we shall define and estimate a function of $\rho(I)$ which is called the index of internal consistency, $\rho(H)$.

Hoyt (ref. 2), was among the first to define an index of internal consistency measure which was derived by variance algebra. In terms of the notation used here, the parametric value of the index is

$$(10)\quad \rho(H) = 1 - V[z(i,1,s)]/V[z(I,1,s)].$$

In equation (10), $V[z(i,1,s)]$ is the variance associated with effects which are not identified with the items of the examination or its administration to the subjects—that is, "error" variance, and $V[z(I,1,s)]$ is the variance associated with the scores of subjects who have taken the examination specified by the conditions identified by the number 1. Thus, if $V[z(i,1,s)]$ nearly equals $V[z(I,1,s)]$, then the index of internal consistency is small. This means that the precision associated with individual scores, composed of responses to I items, is large relative to the "error" variance. In other words, the examination does not discriminate between individuals very well. As a consequence, confidence intervals for individual scores would be relatively large.

It is important to realize the distinction between $\rho(I)$ and $\rho(H)$. The parameter, $\rho(I)$, is a covariance property of the responses to the items of the examination. The parameter, $\rho(H)$, is an index of a relative precision property of a score composed by a simple linear function of the responses to the examination items within a single administration of the examination to each subject. If the examination is administered several times to the same subjects, a derivation similar to the one shown earlier in this paper leads to the evaluation of another covariance property of the scores which is known as the coefficient external consistency. This value is often termed the "reliability" of the examination.

The expressions $V[z(i,1,s)]$ and $V[z(I,1,s)]$ are merely special cases of $\sigma^2[IS]$ and $\sigma^2[S]$ of Table I where we consider only a single method (M=1) or a particular set of conditions—experimental or otherwise. Of course, in this case, $H[IM]$ and $H[M]$ are not tested since no degrees of freedom exist for the tests. Thus the procedure given by Hoyt (ref. 2) is a special case of this analysis.

From Table I,

$$(11)\quad \rho(H) = 1 - \frac{V[z(i,1,s)]}{V[z(I,1,s)]} = 1 - \frac{\sigma^2[1-\rho(I)]}{\sigma^2[1+\overline{I-1}\,\rho(I)]}$$

$$= \frac{I\rho(I)}{1+\overline{I-1}\,\rho(I)} .$$

It is easy to see that the index of internal consistency, $\rho(H)$, is a directly proportional function of $\rho(I)$. Therefore, for a fixed I, as $\rho(I)$ increases, so does $\rho(H)$, and inversely. If we express $\sigma^2[S]$ as a function of $\rho(H)$ rather than $\rho(I)$ we have

$$(12)\quad \sigma^2[S] = I\sigma^2 \,/\, [I - \overline{I-1}\,\rho(H)] .$$

From equation (11), which has the same form as the famous Spearman-Brown prophecy formula, we learn that the index of internal consistency increases as the examination length increases. An increase of the number of covariately consistent items would no doubt increase the variability between subjects. This is shown by $\sigma^2[S] = \sigma^2[1 + \overline{I-1}\,\rho(I)]$. However, the error variance, $\sigma^2[IS] = \sigma^2[1-\rho(I)]$, is invariant to the number of such items. Also, $\sigma^2[IS] = \sigma^2[S]\{1-\rho(H)\}$, which can be shown by simple algebra. Thus we cannot increase the absolute precision associated with the errors of measurement by increasing the examination length, but the relative precision is increased. The parameters $\sigma^2[S]$ and $\sigma^2[IS]$, or their estimates, are often expressed in score units rather than in item units. This is accomplished by multiplying those values by the number of items, I.

A test of the hypothesis that $\rho(I) = 0$, and consequently that $\rho(H) = 0$, is provided by the ratio $SS[S] \,/\, \overline{I-M}\, SS[IS]$ which is distributed as $F[\overline{S-M}, \overline{I-M}\ \overline{S-M}; 1 - \alpha]$. The upper tail of the F distribution should contain the critical region since if $\rho(I) > 0$, the expected value of

TABLE I

ANALYSIS OF VARIANCE FOR THE INTRA- AND INTER-METHOD LINEAR HYPOTHESIS WHEN THE VARIABLES ARE EXAMINATION SCORES

Source of Variation	Degrees of Freedom	Sums of Squares	Expected Mean Squares
$H[I] \; : \zeta(i) = 0$	$\overline{I-1}$	$SS[I] = A-G$	
$H[IM] : \theta(i,m) = 0$	$\overline{I-1}\;\overline{M-1}$	$SS[IS] = B-A-C+G$	
$E[IS]$	$\overline{I-1}\;\overline{S-M}$	$SS[IS] = T-B-D+C$	$\sigma^2[IS] = \sigma^2[1-\rho(I)]$
Intra-Methods Total	$\overline{I-1}\;S$	$SS[TWM] = T-D$	
$H[o] \; : \xi = \xi(o)$	1	$SS[o] = IS\,[y(.,.,.)-\xi(o)]^2$	
$H[M] \; : \mu(m) = 0$	$\overline{M-1}$	$SS[M] = C-G$	
$E[S]$	$\overline{S-M}$	$SS[S] = D-C$	$\sigma^2[S] = \sigma^2[1+\overline{I-1}\,\rho(I)]$
Inter-Methods Total	S	$SS[TBM] = D$	
Grand Total	IS	$SS[T] = T$	

where,

$$A = \sum_i \Big[\sum_m \sum_s y(i,m,s)\Big]^2 / S$$

$$B = \sum_m \sum_i \Big[\sum_s y(i,m,s)\Big]^2 / S(m)$$

$$C = \sum_m \Big[\sum_i \sum_s y(i,m,s)\Big]^2 / IS(m)$$

$$D = \sum_m \sum_s \Big[\sum_i y(i,m,s)\Big]^2 / I$$

$$G = \Big[\sum_i \sum_m \sum_s y(i,m,s)\Big]^2 / IS$$

$$T = \sum_i \sum_m \sum_s \big[y(i,m,s)\big]^2$$

the mean square of $SS[S]$ is greater than the expected value of the mean square of $SS[IS]$. The alternative hypothesis $\rho(I) < 0$ is not of much practical importance because $\sigma^2[S] > 0$ and consequently $\rho(I) > -1/\overline{I-1}$. The $(1-\alpha)\%$ confidence interval for $\rho(I)$ may be obtained from the expression given in (13) wherein $F[o]$ is the ratio of the observed mean squares, $MS[S]/MS[IS]$; $F[U] = F[\overline{S-M}, \overline{I-1}\ \overline{S-M}; 1-\alpha/2]$ and $F[L] = 1/F[\overline{S-M}\ \overline{I-1}, \overline{S-M}; 1-\alpha/2]$ are values obtained from a variance ratio table.

$$(13) \quad \frac{F[o] - F[L]}{\overline{I-1}\,F[L] + F[o]} > \rho(I) > \frac{F[o] - F[U]}{\overline{I-1}\,F[U] + F[o]}$$

A point estimate for $\rho(I)$ can be obtained from (13) by using either the left or right hand side of the inequality and setting $F[L]$ or $F[U]$ equal to 1. The confidence interval for $\rho(H)$ is shown in (14). The point estimate of $\rho(H)$ is found by setting $F[U]$ or $F[L]$ equal to 1.

$$(14) \quad 1 - F[U]\,/\,F[o] > \rho(H) > 1 - F[L]\,/\,F[o].$$

It is also possible, and often desirable, to partition each of the sums of squares and degrees of freedom of Table I into M parts. From these, tests and estimates of $\rho[I(m)]$ and $\rho[H(m)]$ can be made. If we call $r[H(m)]$ the sample estimate of the index of internal consistency for the mth method, then $\rho(H)$, the total index is the weighted average of the $r[H(m)]$. The weights are the sums of squares for the errors, $E[IS(m)]$. Before the tests are carried out by means of Table I, it is advisable to check the hypotheses $\sigma^2[IS(m)] = \sigma^2[IS(m')]$ and $\sigma^2[S(m)] = \sigma^2[S(m')]$ using homogeneity of variance tests.

Summary

Starting from the definition of the linear model of an observation obtained as the result of the administration of an examination in a simple experimental situation, an analysis of variance was obtained whereby tests of hypotheses associated with methods, items, inter-method interactions and the general mean were made. Furthermore, point estimates and confidence intervals were shown for statistics which estimate the:

a. internal consistency—the intra-class correlation of the responses to the items of the examination,

b. index of internal consistency—a measure of the relative precision of a score which is formed by a linear function of the responses to the items of the examination.

The evaluation of the internal consistency of an examination is important for comparison and item construction purposes. The index of internal consistency provides a convenient measure of the relative precision of the error and subject variances of the examination. The derivation given in this paper is easily extended to other common experimental designs.

REFERENCES

1. Cochran, W. G., and Cox, G. M. Experimental Designs (New York: John Wiley and Sons, Inc., 1950).
2. Hoyt, C. J. "Test Reliability Estimated by Analysis of Variance," Psychometrika, VI (1941), pp. 153-160.
3. Nandi, K. H. "A Mathematical Set-Up Leading to Analysis of a Class of Designs," Sankhya, VIII (1947), pp. 172-176.
4. Yates, F. The Design and Analysis of Factorial Experiments, Technical Communication, No. 35 (Harpenden, England: Imperial Bureau of Soil Science, 1937).

CLASSIFICATION BY MULTIVARIATE ANALYSIS WITH OBJECTIVES OF MINIMIZING RISK, MINIMIZING MAXIMUM RISK, AND MINIMIZING PROBABILITY OF MISCLASSIFICATION

M. CLEMENS JOHNSON
University of Minnesota

Principles for Classification

THE PURPOSE of this study is to describe and illustrate certain developments in the field of classificatory analysis which are potentially useful in education. The problem under consideration arises when a number of measurements are made on an individual and it is desired to classify the individual into one of several categories on the basis of these measurements. It is assumed that the individual has been randomly drawn from one of the populations and that in each population there is a probability distribution of the measurements.

In problems of this type it frequently may be observed that the consequences of wrong decisions are not equally undesirable. Somewhat unique is the assumption that it is possible to specify the loss of utility resulting from a wrong decision or misclassification. Loss may be thought of as the penalty paid by the statistician when his guess is incorrect. In this investigation it serves to differentiate the seriousness of the possible errors of classification.

By the risk of committing a certain error in classification is meant the probability of that error multiplied by the corresponding loss. Risk thus corresponds to the expected value of the loss or the expected disutility in the long-run use of some classification rule. In considering the problem of classifying an individual into one of several populations there is a risk or expected loss associated with each possible choice or classification.

Important in the solution of problems of classification are the a priori probabilities whose values postulate the respective chances of drawing an individual from each of the populations under consideration. The criterion by which classification rules are chosen varies with the presence or absence of the a priori probabilities. If they are assumed to be known, rules are selected which minimize overall risk of misclassification. If in a particular problem it is not possible or appropriate to estimate a priori probabilities, the solution requires a different criterion of goodness. Under such conditions it is possible to choose classification rules which minimize the maximum risk of misclassification. This is a conservative approach and represents an application of the minimax principle.

Anderson (1) discusses these criteria in detail. Rao (6), Brown (2), and Welch (9) are useful references. More formal reasoning is generally expressed as follows: Assume, for example, that an individual I with the set of measurements $x_1, \dots, x_q$ has been randomly drawn from one of two populations Π_1 or Π_2 with probability distributions of the measurements $f_1(x_1, \dots, x_q)$ and $f_2(x_1, \dots, x_q)$, respectively. The choice of a classification rule R corresponds to a division of the q-dimensional sample space into two regions R_1 and R_2. If the random point corresponding to the individual falls in R_1 the decision is made that I is from Π_1, and if the point falls in R_2 the decision is made that I is from Π_2.

Let L(2/1) be the loss incurred in classifying I into Π_2 when in fact I belongs to Π_1, and let L(1/2) be the loss incurred in classifying I into Π_1 when in fact I belongs to Π_2. For any rule R let $r_1(R)$ be the conditional risk or expected loss when I belongs to Π_1. It is defined as the probability of I falling in R_2 multiplied by the loss associated with this mistake. Similarly, let $r_2(R)$ be the conditional risk when I belongs to Π_2. It is defined as the probability of falling in R_1 multiplied by the corresponding loss.

Suppose there exist a priori probabilities that I comes from the two populations, say probability p_1 that I does in fact belong to Π_1, and probability p_2 that I belongs to Π_2, where $p_1 + p_2 =$

---

*Summary of unpublished doctoral dissertation, University of Minnesota 1954. Advisor: Palmer O. Johnson, Professor of Education, University of Minnesota.

= 1. It is then possible to define an unconditional risk for any classification rule R as equal to $p_1 r_1(R) + p_2 r_2(R)$. The objective is to find the regions R_1 and R_2 which minimize this risk.

In other problems nothing may be known about the probabilities p_1 and p_2 of drawing from the two populations. Here the unconditional risk cannot be defined and it is necessary to consider the risk $r_1(R)$ when I is from Π_1, and $r_2(R)$ when I is from Π_2. It can be shown that the minimax requirement will be satisfied when $r_1(R) = r_2(R)$.

Multivariate Normal Solutions

Of special interest in application are solutions for the multivariate normal case. Anderson (1) treats these solutions and his results are presented here without accompanying mathematical arguments. For two populations, let $x_1, \ldots, x_q$ have multivariate normal distributions with means in $\Pi_1 = u_i^1$ and in $\Pi_2 = u_i^2$, $i = 1, \ldots, q$. Assume the two distributions have common variances and correlations. The regions for minimum risk are defined by the following discriminant functions or classification rules:

$$R_1 : \sum_{i=1}^{q} \lambda_i x_i - 1/2 \sum_{i=1}^{q} \lambda_i (u_i^1 + u_i^2) \; 7 \; \log_e k$$

$$R_2 : \sum_{i=1}^{q} \lambda_i x_i - 1/2 \sum_{i=1}^{q} \lambda_i (u_i^1 + u_i^2) \angle \log_e k$$

where $\lambda_1, \ldots, \lambda_q$ is the solution of

$$\sum_{j=1}^{q} \sigma_i \sigma_j p_{ij} \lambda_j = u_i^1 - u_i^2 \text{ and}$$

$$k = \frac{L(1/2)p_2}{L(2/1)p_1}$$

If $L(1/2) = L(2/1)$, $k = p_2/p_1$ which is the solution of minimum probability of misclassification. If a priori probabilities are not specified, $\log_e k(=c)$ may be found so that the risk when I is from Π_1 is equal to the risk when I is from Π_2. The solution employs the univariate normal table.

Similar formal solutions are obtained for m multivariate normal populations with the same set of variances and correlations. In practice these solutions present great difficulties when the number of populations is greater than two. For example, when three groups are considered it is necessary to assume equal costs of misclassification.

Assumptions

Of basic importance is the assumption that it is possible to assign a comparable numerical value to each of the losses incurred in misclassification. While presently this assumption is difficult if not impossible to satisfy with rigor, it should be noted that the usual assumption of equal costs of misclassification is often unrealistic. Utility measurement is an important contemporary problem in the economic field. The reader is referred to Friedman and Savage (4) for one account of the historical background of utility discussions. An undecided issue appears to be whether utility (or disutility) is actually measurable or whether any choice among alternatives depends merely on their ordering. Coombs (3) in the social sciences suggested the use of a scale not involving a unit of measurement designed to order the magnitude of the interval between objects. The scale was in addition to the types set up by Stevens (7) and fell logically between the interval scale and the ordinal scale.

It is assumed in the model that the parameter values in the multivariate normal distributions are known exactly. This is seldom true in practical applications. Although Wald (8) and Anderson(1) have developed classification statistics which employ sample estimates for parameter values in problems involving two multivariate normal populations with the same covariance matrix, the distributions are presently unsuitable for practical use. A reasonable solution consists in obtaining the best possible estimates and substituting them for the unknown parameters in obtaining the classification rules. For sufficiently large samples it is possible to proceed as if the parameters were known though some unknown error is thereby introduced.

Illustrative Problem

The General College of the University of Minnesota maintains a staff of trained counselors who devote their time primarily to helping individual students. A principal function of these counselors is to assist first quarter students in the selection and planning of an appropriate educational goal. In this selection process it is important to consider the capabilities of the new student with respect to other groups of General College students. Do his capabilities most nearly resemble those of students who drop out of the General College, students who are accepted for transfer by another college of the University, or students who continue in the General College to receive an associate in arts degree? These are the groups employed in classification.

Recommendations are customarily made on a judgmental basis using information available to the counselor and the beginning student. Objective measures frequently used for this purpose include high school percentile rank, score on the American Council on Education Psychological Examination, score on the Cooperative English Test, and scores on the Comprehensive Examination of the General College. These are the variables employed for establishing the classification rules.

It was considered desirable in the design of this investigation to provide for comparable groups on which to establish classification rules and cross validate the results. This was done by selecting a random half of all students matriculating in the General College during the fall quarters of 1949 and 1950. Within each half students defined by population restrictions were assumed to be samples of their respective populations. Sample sizes in the main analysis were 273 for the drop out group, 76 for the graduate group, and 79 for the transfer group. Sample sizes used in cross validation were 264 for the drop out group, 75 for the graduate group, and 88 for the transfer group.

Classification Into One of Two Groups

Applications were made to the problem of classifying students into one of two categories: (a) students who were accepted for transfer by another college of the University and did not graduate from the General College and (b) students who graduated from the General College. Similar classification rules were established using two different sets of variables. The duplication was made to examine the models under different conditions of application. For each set of variables three solutions were obtained corresponding to the following standards of goodness: minimizing risk of misclassification, minimizing maximum risk of misclassification, and minimizing probability of misclassification. Mean values for the two groups on four selected variables were as follows:

	Transfer Π_1	Graduate Π_2
ACE	94.16	87.83
Eng.	167.38	162.62
HSR*	4.51	4.41
Comp.	60.28	53.43

*profit values

The dispersion matrix based on the two groups is shown below:

$$\begin{vmatrix} 248.49 & 251.30 & 1.06 & 110.70 \\ 251.30 & 930.73 & 1.90 & 167.45 \\ 1.06 & 1.90 & .39 & 1.46 \\ 110.70 & 167.45 & 1.46 & 216.47 \end{vmatrix}$$

The regions were defined by the following rules or discriminant functions:

$$R_1: .0185X_1 - .0046X_2 + .1494X_3 + .0247X_4 - 2.995 > \log_e k$$

$$R_2: .0185X_1 - .0046X_2 + .1494X_3 + .0247X_4 - 2.995 < \log_e k$$

where values of k were chosen according to the described criteria of goodness.

Approximations of losses due to misclassification were obtained from three experienced counselors in the General College. Since the counselors were not in agreement, a median judgment was employed in the analysis where disutility in misclassifying transfers = 3 and disutility in misclassifying graduates = 2. Values assumed for the a priori probabilities were based on the relative frequencies of the two groups. These values were .51 and .49 for transfers and graduates, respectively. Subsequently, the solution of minimum risk correctly classified a large majority of students actually belonging to the transfer group, whereas the same solution misclassified an equally large majority of students belonging to the graduate group. The tendency to classify most people as transfers was pronounced since the difference between groups was not very large while the difference in disutilities was comparatively great. A similar pattern should result if a counselor were to employ the same variables under a criterion of minimum risk.

The solution in which maximum risk of misclassification was minimized produced classification frequencies which were intermediate to those obtained by the other solutions. This rule was not as efficient in reducing overall risk since it was based on a lesser amount of information, namely, the a priori probabilities of drawing from the two populations. In terms of overall number of correct classifications the "best" solution was that in which probability of misclassification was minimized. This solution assumed equal costs of misclassification and, consequently, represented a different criterion of goodness. Tables I, II and III present the results of the cross validation studies for the various classification rules.

Classification Into One of Three Groups

In the three group solutions it was necessary for purposes of technical simplification to assume equal costs of misclassification. This essentially meant that classification rules were chosen on the basis of probabilities of misclassification rather than risks of misclassification. Another distinguishing feature of the three group solution was the use of the bivariate normal table (5) to obtain the minimax solution.

Students were assigned to one of three groups: (a) transfer, (b) graduate, and (c) drop out. Separate rules were established with objectives of minimizing probability of misclassification and

TABLE I

CLASSIFICATION BY SOLUTION OF MINIMUM RISK

True Category	Predicted Category			Probabilities of Misclassification	
	Transfer	Graduate	Total	Empirical	Predicted
Transfer	73	15	88	.17	.14
Graduate	63	12	75	.84	.71
Total	136	27	163	.50	.42

TABLE II

CLASSIFICATION BY MINIMIZING THE MAXIMUM RISK OF MISCLASSIFICATION

True Category	Predicted Category			Probabilities of Misclassification	
	Transfer	Graduate	Total	Empirical	Predicted
Transfer	54	34	88	.39	.32
Graduate	46	29	75	.61	.48
Total	100	63	163		

TABLE III

CLASSIFICATION BY SOLUTION OF MINIMUM PROBABILITY OF MISCLASSIFICATION

True Category	Predicted Category			Probabilities of Misclassification	
	Transfer	Graduate	Total	Empirical	Predicted
Transfer	50	38	88	.43	.37
Graduate	37	38	75	.49	.42
Total	87	76	163	.46	.39

TABLE IV

CLASSIFICATION BY SOLUTION OF MINIMUM PROBABILITY OF MISCLASSIFICATION

True Category	Predicted Category				Probabilities of Misclassification	
	Transfer	Graduate	Drop Out	Total	Empirical	Predicted
Transfer	15	0	30	45	.66	.79
Graduate	3	0	31	34	1.00	.99
Drop Out	4	0	117	121	.03	.04
Total	22	0	178	200	.32	.35

TABLE V

CLASSIFICATION BY MINIMIZING THE MAXIMUM PROBABILITY OF MISCLASSIFICATION

True Category	Predicted Category				Probabilities of Misclassification	
	Transfer	Graduate	Drop Out	Total	Empirical	Predicted
Transfer	20	19	6	45	.56	.55
Graduate	7	21	6	34	.38	.55
Drop Out	14	50	57	121	.53	.55
Total	41	90	69	200		

minimizing maximum probability of misclassification. Customarily, only one standard of goodness would be adopted depending upon the presence or absence of a priori probabilities of drawing from the three populations. In this example probability of drawing from the transfer group = .18, from the graduate group = .18, and from the drop out group = .64.

In the solution of minimum probability of misclassification most students in the cross validation groups were assigned to the drop out category. This was because differences between groups were not very great, while differences between a priori probabilities were comparatively large. When probabilities were not assumed a solution was obtained which minimized the maximum probability of misclassification. Although not as efficient in terms of overall number of corect classifications, the minimax solution classified students in a more normal fashion. Tables IV and V summarize the cross validation results for the three group solutions based on a sample of 200 students drawn at random from all validation groups.

Summary

This investigation had its setting in the General College of the University of Minnesota and illustrated the classification of students into educational groups on the basis of measures commonly used for counseling purposes. The groups were assumed to represent multivariate normal populations with the same set of variances and correlations. Comparable samples were obtained on which to establish classification rules and cross validate results. This was done by selecting a random half of all students matriculating in the General College during the fall quarters of 1949 and 1950.

Principles for choosing a rule of classification were based on an approximation of the losses incurred with misclassification. By the risk of committing an error in classification was meant the probability of that error multiplied by the corresponding loss. The criterion by which a rule was chosen varied with the presence or absence of a priori probabilities of drawing an individual from the populations. If probabilities were assumed to be known, a rule was selected which minimized risk of misclassification. Assuming it was not possible or appropriate to assign a priori probabilities, a rule was chosen which minimized the maximum risk of misclassification. When three groups were involved it was necessary to assume equal costs of misclassification.

The techniques are applicable in theory to most classification problems. They should be useful, for example, in making decisions or whether to admit or refuse to admit students to college. While a rigorous determination of utility values remains a serious obstacle, an approximated value may sometimes be more realistic than the assumption of no differences.

REFERENCES

1. Anderson, T. W. ‘‘Classification by Multivariate Analysis'', Psychometrika, 16 (1951), pp. 31-50.
2. Brown, George W. ‘‘Basic Principles for Construction and Application of Discriminators'', Journal of Clinical Psychology, 6 (1950), pp. 58-61.
3. Coombs, C. H. ‘‘Psychological Scaling Without a Unit of Measurement'', Psychological Review, 56 (1950), pp. 145-158.
4. Friedman, Milton and Savage, L. J. ‘‘The Utility Analysis of Choices Involving Risk'', Journal of Political Economy, 57 (1948), pp. 279-304.
5. Pearson, Karl. Tables for Statisticians and and Biometricians, Part II (Cambridge: Cambridge University Press, 1931).
6. Rao, C. R.‘‘Utilization of Multiple Measurements in Problems of Biological Classifications,'' Journal of the Royal Statistical Society, B 10 (1948), pp. 159-203.
7. Stevens, S. S.‘‘On the Theory of Scales of Measment,'' Science, 103 (1946, pp. 677-680.
8. Wald, Abraham. ‘‘On a Statistical Problem Arising in the Classification of an Individual into One of Two Groups'', Annals of Mathematical Statistics, 15 (1947), pp. 145-163.
9. Welch, B. L. ‘‘Notes on Discriminant Functions'', Biometrika, 31 (1939), pp. 218-20.

THE POISSON DISTRIBUTION IN EDUCATIONAL RESEARCH

SAM DUKER
Brooklyn College

EDUCATIONAL research in many instances consists of gathering data about a sample and then generalizing the resulting findings to the population from which the sample was selected. The inferences thus made are valid only if, first, the sample is properly selected, and second, appropriate statistical techniques are employed. The interesting and often quite involved questions involved in the process of sampling are not within the scope of this paper. Here only the phase of the statistical process by means of which inferences are made about a defined population on the basis of data which are collected from a sample of that population is dealt with.

Fundamentally, such inferences can only be drawn when the shape of the distribution formed by the variables making up the population can be justifiably hypothesized. In educational research there has been a noticeable tendency to assume that all distributions are of the normal or bell-shaped curve type. This assumption is not always justified. The normal distribution is a very convenient statistical model; it is neatly symmetrical and mean, mode and median coincide. In real situations we seldom find such neat symmetry nor do we find many cases where the mean, mode and median coincide. These statments in no way depreciate the importance of the concept of the normal distribution for this concept is an extremely important and useful one in educational research.

When the shape of the distribution followed by the population is not known it becomes necessary to use non-parametric statistical devices. Such techniques are applicable regardless of the shape of the distribution and are therefore highly useful. More accurate and usable results are possible, however, when the shape of the distribution can be validly hypothesized.

Many distributions other than the normal are known to statisticians. One of these is the Poisson distribution. Curiously enough, the existence of this distribution was first noted in 1740 by Abraham de Moivre in his *Doctrine of Chances* (1) and later by the man after whom it is named. Simeon Denis Poisson, a French mathematician who lived from 1781 to 1840 and who was a man of rare diversity of talents, developed a derivation for this distribution in 1820 (2) in a work on the application of the laws of probability to the determination of the likelihood of miscarriages of justice in courts of law. Since that time this distribution has been independently derived by a number of other investigators. The Poisson distribution may be derived either as a limit of the binomial distribution or as an independent distribution of randomly distributed rare events.

The Poisson distribution is a discrete function which exists only for positive integral values of the variate. When the mean is less than one the probability polygon is J-shaped. When the mean is greater than one it becomes double sided and for larger values of the mean tends toward symmetry. This is another way of saying that when the mean is less than one, the mode is zero. When the mean is greater than one, but not an integer, the mode is the nearest integer less than the mean. When the mean is an integer greater than or equal to one the probability polygon is bimodal, the modes being the mean and the integer immediately below the mean.

In the Poisson distribution the mean (which is equal to the variance) performs a function similar to that performed by the mean and the standard deviation together in the normal distribution, for knowledge of the mean enables one to determine the distribution and all its moments.

The Poisson distribution is rarely employed in educational research. The ever increasing use of statistical techniques in educational research makes it important that other than normal distributions be employed where necessary. Not all non-normal distributions will follow the Poisson law, but, when the assumptions underlying the Poisson distribution appear to be met, the hypothesis that the data may be distributed in such a manner should be tested. More accurate analysis and consequently more reliable statistical inferences are possible when the true nature of a distribution is recognized, whatever its form may be.

Data may be expected to fall into the form of the Poisson distribution only when certain basic assumptions are met. These assumptions are five in number.

1. *Discreteness*. The data involved must be discrete rather than continuous. The event, the occurrence of which constitutes the data, must be of the type which either occurs or does

not occur. It is an "all or none" proposition. Scores obtained from measurements yielding fractional or negative values cannot, therefore, under any circumstances, be said to constitute data subject to the Poisson law.

2. Possibility of large number of occurrences of the event. The possible number of occurrences of the events at any particular opportunity must theoretically be infinite. In practice, of course, such data are not to be expected. It is essential, however, that in such specified division of time, space or other unit, there must exist a possibility of a large number of occurrences of the event, the frequency of the occurences of which constitutes the variable making up the distribution in question. This requirement must not be confused with a statement concerning the size of the sample under consideration. The size of the sample could not influence the nature of the population from which the sample is drawn. If, however, the sample is small the form of the population distribution cannot be determined with much confidence.

When the possible number of occurrences is small, a distribution is often observed to have frequencies conforming closely to the theoretical frequencies for zero, one and two. The hypothesis that the Poisson distribution is present must be rejected in such cases, as it must in all cases where the basic assumptions are not met.

3. Small likelihood of the event occurring at any particular opportunity. The hypothesis that a Poisson distribution is present cannot be accepted unless the probability that the event will occur at any one given opportunity is very small. The requirement that there be a large number of possible occurrences must be distinguished from the present limitation. It must be borne in mind that in order to establish the presence of a Poisson distribution both of these requirements must be met and that compliance with only one of them is insufficient.

4. Equality of opportunity for occurrence. The specified divisions of time, space, or other units which are so constituted that the number of occurrences of an event or the frequency of individuals can be counted for each division present "areas of opportunity for occurrences." Each such division or area must present an equal opportunity for the occurrence of the event being observed. It would not be possible, for example, to use as such an area or division, schools of a city or state when because of varying qualities as to size or excellence, such schools present varying probabilities of an event occurring or not occurring in a division. In other words, where the probability of the occurrence of an event varies from division to division, whether such division be one of time, place, or other unit, the Poisson distribution may not be used as a model for the empirical distribution. Sometimes a compound Poisson distribution may occur in such cases.

5. Independence. The occurrences of the event must be independent, that is to say, a. the occurrences of an event at one opportunity must have no effect on the probability of a second occurrence of the event at the same opportunity, and b. the occurrence of the event at one opportunity must have no effect on the probability of an occurrence of the event at another opportunity.

The assumption of independence is of great importance in determining the propriety of accepting a hypothesis that the Poisson distribution is present. It is essential that the occurrence of that event at that same time or place or at another time or place. Many educational data seem to conform to the Poisson distribution in all other respects such as equality of mean and variance, correspondence of observed and theoretical Poisson frequencies and the meeting of all other assumptions, but fail to meet this requirement.

There are, for example, several situations having to do with new processes in education in which at first glance the distributions of the frequency of adoption of such new processes appear to follow the Poisson law. Careful analysis, however, reveals that the adoption of such a process in one situation may well tend to influence its adoption in another and that, therefore, the requirement of independence is not fulfilled. Under such circumstances accepting the hypothesis that the Poisson distribution may be regarded as a model would be erroneous.

In general, the Poisson distribution will be found only in cases where the distribution is highly skewed, but this is not necessarily always so. As already noted, the Poisson distribution tends toward symmetry when the mean is large.

In a Poisson distribution the mean and variance are equal. While the near equality of mean and variance certainly is a rough indication that the possibility of the presence of a Poisson distribution should be investigated, it is not a sufficient basis for assuming the presence of such a distribution. Distributions with equal or nearly equal mean and variance may fail to meet the assumptions underlying the Poisson distribution. In such cases the hypothesis that this distribution is followed by the data must be rejected even when the observed frequencies equal or approach the tabulated frequencies of the Poisson distribution.

It is important to know that an empirical distribution of educational data follows the Poisson law because with such knowledge a better analysis and understanding of the data are possible. First, it is possible to estimate the liklihood of an event occurring or not occurring with greater accuracy because of the availability of tables of the Poisson distribution. Second, it is vital to the use of various statistical tests to know the nature of the distribution involved because such tests are valid only when certain basic assumptions underlying them have been met. When the basic assumptions for such a test are not known to be satisfied by data distributed in a Poisson distribution the use of such a test cannot be justified. Third, it is important in performing an analysis of variance, as a transformation is known which makes possible the application of this procedure to data distributed in accordance with the Poisson law. It is here that the educational research worker is particularly concerned because he so frequently uses analysis of variance procedures. Often such an analysis is performed without any examination of the presence or absence of equality of variability in the distribution of the data being examined. Since such equality of variability is a basic assumption underlying this technique, its absence is fatal to the correctness and validity of the results obtained. When a difference is claimed to be significant or non-significant on the basis of a very small departure from a given F-value, such a failure to obtain precise results may invalidate an important conclusion.

The number of instances in which the Poisson distribution has been found to be applicable in research fields other than education is very large. Only a few examples can be given here.

Research workers in the field of accident proneness have found the Poisson distribution to be fundamental to their work (3).

Many situations occurring in both pedestrian and vehicular traffic have been found to fit the Poisson form (4). This fact has been utilized in tabulating the most efficient timing of traffic lights as well as the most efficient placing of highway stop signs.

Many natural phenomena occur in such a random manner as to meet all the assumptions underlying the Poisson distribution. A few of them are radio-active disintegration (5), chromosome breakages produced in cells by X-Ray (6), the number of nuclei of condensation in the air (7), and the number of meteoric falls per small geographic region (8).

Quality control tables for sequential sampling purposes are based on this law (9).

The new field of operations research has been developed in large part on the basis of the Poisson law (10).

Telephonic engineers use the Poisson distribution in determining the number of cables, switchboard stations, and so forth needed to serve a given amount of traffic (11).

In concluding this abbreviated list of non-educational situations involving this distribution, a few odd instances will be given.

Singh (12) found that the daily number of ticketless passengers traveling on an Indian railway were distributed normally but that the number who had to be evicted for refusal to pay their fare fell into the form of a Poisson distribution.

The classic example is the one given by Bortkewitz (13) who found that the number of men per year per Prussian army corps who were killed by the kick of a horse fell into this form.

In conclusion one example of educational data distributed in accordance with the Poisson law will be discussed.

When a rote task of the sort in which perfect mastery is the goal is first introduced to learners, the probability of errors being made by these learners is very large. This probability becomes smaller as the stage of perfect mastery of the rote task is approached. Of course, the possibility of making errors remains equally large at all stages of the learning process. It is also reasonable to conclude that at the beginning of the instruction, the commission of one error increases the probability of another error. In typewriting, for example, a beginner hitting a wrong key is likely to repeat that error or to make other errors based on his erroneous mental picture of the keyboard. An experienced student is not so likely to make cumulative errors. This is another way of saying that the errors made by advanced learners are more nearly independent than the errors made by beginners. In this type of learning, the possible number of errors in one task is the same as that in another. It, therefore, appears that a distribution of errors made by learners, as the stage of perfect mastery is approached, meets all the requirements imposed by the basic assumptions underlying the Poisson distribution. If this is so, some interesting implications arise concerning a method of measuring the state of learning attained by a group being instructed in the type of task here discussed. When the distribution of errors takes the form of a Poisson distribution it seems reasonable to suggest the hypothesis that the point has been reached in the learning process where further instruction would yield minimal further gains in learning. A further hypothesis that such a distribution would depart from the Poisson form when instruction and practice cease and forgetting sets in, would seem worthy of investigation. The hypothesis that further practice would then again bring such a distribution into the shape of the Poisson distribution would follow logically if the first two hypotheses were found to be tenable.

The relationship between the distribution of errors and the Poisson distribution was first pointed out by Shuster (14) in his study on the teaching of slide rule procedures. Shuster found that the number of errors made by his students at the end of the experiment were distributed in accordance with the Poisson law.

With this study as a cue a study was made of the errors made by students of typewriting (15).

The subjects were students taking first, second, and third semester courses in typewriting in the Commerce Department of a large midwestern college. In each of these courses a bi-weekly test of constant length was given. The data covered a period of three years and included the results of nearly 9000 of these bi-weekly tests. An examination of the average number of errors made per student per test showed that there was little change in this respect (the mean varied from a high of 8.98 in the first semester to a low of 6.29 in the last test of the third semester). This fact is easily accounted for by the increased rates of speed on the part of the more advanced students. There was, however, a considerable change in the variance which dropped from 36.47 in the first test of the first semester to 7.15 in the last test of the third semester.

The distribution of errors in the third semester showed a clear correspondence to the Poisson distribution and the chi-square test bore out this correspondence.

A specific examination of the assumptions heretofore given as applied to this situation bears out the reasonableness of this correspondence between the empirical distribution and the theoretical Poisson distribution.

1. Discreteness. The distributions are discrete as an error is either made or it is not. There is no other possibility. Negative or fractional values are impossible in the situation considered here.

2. Possibility of a large number of occurrences of the event. There are a very large number of possibilities for errors to be made in each test and this requirement is, therefore, rather obviously met.

3. Small likelihood of the event occurring at any particular opportunity. As perfect mastery is approached by the student, the probability of an error at any particular single opportunity becomes constantly smaller. It would, therefore, appear that the requirement that the probability of an event at any particular instance should be small will be fulfilled by distributions of errors made by learners who are approaching the final stages of learning.

4. Equality of opportunity for occurrence. Very little question could arise to cast doubt on the theory that each of the many items in the test present an equal opportunity for the occurrence of the event we are concerned with.

5. Independence. As has already been noted, the kinds of errors made at the later stages of learning of a rote process are not dependent on other errors as errors at the early stages may be. The commission of an error by an advanced student does not make appreciably more likely the commission of another error by him or by any other student. The requirement of complete independence of the events with the frequency of which the distributions are concerned is, therefore, substantially complied with at later stages of instruction.

In the light of this examination of the manner in which these distributions meet the basic assumptions underlying the Poisson distribution it would appear reasonable that the frequency of the errors made by advanced students would fall into the form of the Poisson distribution while the earlier distributions do not.

A similar study is reported by Davis (16) in which he found that data consisting of the number of typing strokes between errors fell into this form.

Many other examples of Poisson distributions in the field of educational research could be given if space permitted.

While preparing this paper the writer came across the following sentence in an article purporting to report an experiment in the teaching of reading: "With no emphasis upon the formal teaching reading whatever, the experimental groups outgained the comparison groups in reading achievement by an almost double score (9.5 months to 5.8 months) at a statistical significance of less than 1%." It is hoped that this discussion of the Poisson distribution may contribute to reports of educational research having a statistical significance of more than 1%.

REFERENCES

1. Moivre, Abraham de. The Doctrine of Chances or a Method of Calculating the Probability of Events in Play. London, W. Pearson, 1718.
2. Poisson, Simeon Denis. Recherches sur la Probabilite des Jugemens en Matiere Criminelle et en Matiere Civile Precedees des Regles Generales du Calcul des Probabilites. Paris, 1837.
3. Greenwood, Major and G. Udny Yule. "An Inquiry into the Nature of Frequency Distributions Representative of Multiple Happenings with Particular Reference to the Occurrence of Multiple Attacks of Disease or of Repeated Accidents." Journal of the Royal Statistical Society, 83:255-79, 1920.
4. Greenshields, Bruce D., Donald Schapiro, and Elroy Erickson. Traffic Performance at Urban Street Intersections. Yale Bureau of Highway Traffic, Report No. 1. New Haven, Conn., Yale University, 1947.
5. Rutherford, Ernest and H. Geiger. "The Probability of Variates in the Distribution of Alpha-Particles." London, Edinburgh, and Dublin Philosophical Magazine and

Journal of Science, 20:698-707, 1910.

6. Feller, William. An Introduction to Probability Theory and Its Applications. New York, John Wiley and Sons, Inc., 1950.
7. Scrase, F. J. "The Sampling Error of the Aitken Nucleus Counter." Quarterly Journal of the Royal Metereological Society, 61:367-79, 1935.
8. Hoel, Paul G. "On the Probability of Coincidental Metereological Falls by Means of Poisson Frequencies." Popular Astronomy, 55:45-48, January, 1947.
9. Dodge, H. F. and H. G. Romig. "Single Sampling and Double Sampling Inspection Tables." The Bell System Technical Journal, 20:1-61, January 1941.
10. Morse, Phillip M. and George E. Kimball. Methods of Operations Research. Boston, The Technology Press of Massachusetts Institute of Technology, 1951.
11. Thorndike, Frances. "Applications of Poisson's Probability Summation." Bell System Technical Journal, 5:604-24, Jan. 1926.
12. Singh, J. "Incidence of Ticketless Travelling on the Luchnow Campore Section of the East Indian Railway." Sankhya, 7:211-12, November 1945.
13. Bortkewitsch, L. von. Das Gesetz Der Kleinen Zahlen. Leipzig, B. G. Teubner, 1898.
14. Shuster, Carl N. A Study of the Problems of Teaching the Slide Rule. Teachers College Contributions to Education, No. 805. Bureau of Publications, 1940.
15. Duker, Sam. "The Poisson Distribution and Its Significance to Educational Research." Unpublished doctoral thesis, Columbia University, 1952. Typewritten.
16. Davis, D. J. "An Analysis of Some Failure Data." Journal of the American Statistical Association, 47:113-50, 1952.

FORECASTING FUTURE ENROLLMENT BY CURVE-FITTING TECHNIQUES

H. P. KUANG*
University of Minnesota

WE KNOW relatively little in advance about the future, yet we are frequently making plans for the future. Every university president or every college dean is constantly interested in knowing how many dollars, buildings, numbers of the faculty and course offerings would be required to provide instruction for the students who would be in attendance in one or in a series of years in the future. This requires in a high degree the forecast of future college enrollment.

There are a number of approaches which have been made to the problem of forecasting school enrollment. The approach as discussed here involves a curve-fitting technique. The curve-fitting procedure consists in determining a functional relationship which exists between years and the past enrollment records and then projecting this function to the year or years for which prediction is desired. It should be noted that this approach is made on the fundamental assumption that the historical records provide a useful clue to the possible future enrollments. In other words, it is assumed that the present enrollment growth is fairly well indicated by past records, and that factors operating in the recent past will continue their influences in the future.

The purpose of this paper is to suggest three types of curves by means of which the prediction of future enrollment might be made. It is not my attempt to review all growth curves, but merely to give a little suggestion on the problem of estimating uncertain enrollments.

A. The Exponential Curve

The exponential function is applicable to the growth of enrollment. We shall first define the function as follows:

If the values of t (say, year) are in arithmetic progression and the corresponding values of y (say, enrollment) are in geometric progression, then the relation between the variables t and y is said to be an exponential function, $y = ab^t$.

Let t_1, $t_1 + \Delta t$, $t_1 + 2\Delta t$ .... $t_1 + (n-1)\Delta t$ be an arithmetic series and let y_1, ry_1, r^2y_1 .... $r^{n-1}y_1$ be a geometric series. Then the t_nth term of the arithmetic series is $t_n = t_1 + (n-1)\Delta t$ or $n-1 = \frac{t_n - t_1}{\Delta t}$. Similarly, the y_nth term of the geometric series is $y_n = r^{n-1}y_1$.

$$\text{Hence} \quad y_n = y_1 r^{\frac{t_n - t_1}{\Delta t}} = y_1 r^{\frac{-t_1}{\Delta t}} (r^{\frac{1}{\Delta t}})^{t_n}$$

$$\text{or} \quad y_n = ab^{t_n} \tag{1}$$

$$\text{where} \quad a = y_1 r^{\frac{-t}{\Delta t}} \quad \text{and} \quad b = r^{\frac{1}{\Delta t}}$$

Since any t is connected with the corresponding y the equation (1) becomes

$$y = ab^t \tag{2}$$

The exponential curve may be fitted readily by reducing it to logarithmic form. Taking logarithms of the above expression (2), the equation becomes

$$\log y = \log a + t \log b \tag{3}$$

which is a straight-line trend of log values.

According to the least-square principle we may minimize the residuals by differentiating in turn with respect to log a and log b. We then obtain the solution of the logarithms exponential equation (3):

$$\log a = \frac{\Sigma t^2 \Sigma \log y - \Sigma t \Sigma t \log y}{n\Sigma t^2 - (\Sigma t)^2}$$

$$\log b = \frac{n\Sigma t \log y - \Sigma t \Sigma \log y}{n\Sigma t^2 - (\Sigma t)^2}$$

In case the data are not given with the values of the independent variable in arithmetic progression it may be readily proved that if the ratio of the first difference of log y to the first difference of t is constant, then the relation between t and y can still be expressed by the same exponential function (2).

---

*The author is grateful to Professor Palmer O. Johnson, who read critically the draft of this paper. Thanks are also due to Dean R. E. Summers through whose courtesy the author was permitted to spend a short period of time examining the trend of the enrollment data in his office.

B. The Gompertz Curve

This curve was first reported in 1825 in a paper read by Benjamin Gompertz, a Jewish statistician who worked with a life insurance company in England. The general formula for the curve is

$$y = d \cdot g^{q^t} \tag{4}$$

in which y represents measure of growth at the time t; d, g and q being constants. Essentially the Gompertz curve is of the exponential type. The exponential function as discussed above is a linear trend of log values while the Gompertz curve is a linear trend of log log values. To illustrate this relationship we simply put $d = 1$ and take the logarithm of both sides of the equation (4)

$$\log y = q^t \log g$$

Taking the logarithm again, we have

$$\log \log y = t \log q + \log \log g$$

which is the equation of a straight line in terms of log log values.

Hence, it is evident that the Gompertz function is expression of the nature of enrollment growth if the relationship between time and the log log values of the enrollment is linear. In order to fit the enrollment data to the Gompertz curve, we have to determine the constants d, g and q. The three constants may be estimated from three equally spaced points which are so chosen as to represent the characteristic of the data. Let the three points be $(t_1, y_0)(t_1 + 1, y_1)$ $(t_1 + 2, y_2)$. Substituting in the equation (4), we have the following three equations:

$$y_0 = d \cdot g^{q^{t_1}}$$

$$y_1 = d \cdot g^{q^{t_1+1}}$$

$$y_2 = d \cdot g^{q^{t_1+2}}$$

To find values for d, g and q in terms of y_0, y_1 and y_2 we take the logarithm of both sides of each of the three equations and solve them. The value for d, g and q for fitting the Gompertz curve are as follows:

$$\log d = \frac{q \log y_0 - \log y_1}{q - 1}$$

$$\log g = \frac{\log y_1 - \log y_0}{qt(q-1)}$$

$$q = \frac{\log y_2 - \log y_1}{\log y_1 - \log y_0}$$

C. The Logistic Curve

In 1838 Professor P. F. Verhulst, a Belgian mathematician, published a memoir suggesting the use of a curve, which he called the logistic, to describe the growth of population. His work was forgotten for seventy years, but since the publication of Du Pasquier's paper in 1918, the logistic curve has been widely used in the study of population gorwth. Since there is a marked relation between the past growth of population and the growth of enrollment, it is not novel to apply the logistic curve in the fitting of enrollment trend.

The general logistic function may be given by

$$y = \frac{L}{1 + me^{\Phi(t)}} \tag{5}$$

where $\Phi(t) = S_1 t + S_2 t^2 + S_3 t^3 + \dots\dots\dots + S_n t^n$ y denotes the growing variable, enrollment, t denotes time, year; m, L and S's are constants.

Several methods have been developed for fitfing this curve. The first of these, which may be called the method of the rate of increase, is based upon the fundamental logistic property that the percentage of rate of increase is a linear function of the growing variable, say enrollment. We may calculate the percentage of increases over successive intervals and plot these to the enrollments at the end of the intervals, then, fitting a line to these points, say by least squares. The application of this procedure is given by H. Hotelling. Another method, which may be called the method of selected points, is used by Verhulst and by Pearl and Reed. This method consists essentially of a preliminary estimate of the constants or parameters from a set of points. The Pearl-Reed procedure for estimating the constants is limited to five parameters. It should be pointed out, however, that any number of parameters may be determined by a series of points. Suppose we wish to determine n+2 parameters by use of n+2 points $(0, y_0)(t_1, y_1)(t_2, y_2)$ $\dots\dots\dots (t_n, y_n)(t_{n+1}, y_{n+1})$, we have from (5):

$$me^{\Phi(t)} = \frac{L - y}{y}$$

$$\Phi(t) = \log (L - y) - \log m - \log y = \log \frac{L - y}{y} - \log m$$

$$\text{Since } \Phi(t) = S_1 t + S_2 t^2 + \dots\dots\dots\dots + S_n t^n$$

Let $\log m = S_0$
Then $S_0 + S_1 t + S_2 t^2 + \dots\dots\dots + S_n t^n =$

$$\log \frac{L - y}{y}$$

Now if the selected n+2 points pass through the curve we can determine n+2 parameters, S_0, S_1, $S_2 \dots\dots S_n$ and L by the following equations:

$$S_0 - \log \frac{L - y_0}{y_0} = 0$$

$$S_0 + t_1S_1 + t_1^2S_2 + \dots\dots\dots + t_1^nS_n - \log\frac{L-y_1}{y_1} = 0$$

$$S_0 + t_3S_1 + t_2^2S_2 + \dots\dots\dots + t_2^nS_n - \log\frac{L-y_2}{y_2} = 0$$

$$\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots$$

$$S_0 + t_nS_1 + t_n^2S_2 + \dots\dots\dots + t_n^nS_n - \log\frac{L-y_n}{y_n} = 0$$

$$S_0 + t_{n+1}S_1 + t_{n+1}^2S_2 + \dots\dots + t_{n+1}^nS_n - \log\frac{L-y_{n+1}}{y_{n+1}} = 0$$

The values of these parameters may be denoted in terms of determinants or can be obtained by any of the algebraic methods for solving simultaneous equations, such as the Doolittle method. The solution of these equations can be treated as a first approximation to the best fitting logistic curve. If a better fit is desired than that given by these selected points, we may use the second approximation by minimizing the sum of squares of the residuals; and the process may be repeated a number of times. This process does not, however, minimize the sum of the squares of the differences between the observed and expected values. Hence, strictly speaking, the procedure, as called by Pearl and Reed the method of least square, is not true least square procedure in the usual sense. The details of the least square fitting are given by Henry Schultz (ref. 4).

A good many statisticians and mathematicians in a good many lands have attempted to devise a curve for the growth phenomenon. Mitscherlich of Germany, Robertson of Australia, Yule of England have all used a type of curve to define growth. There is no demonstration of the applicability of these curves, so far as I am aware, to the enrollment growth.

In projecting future enrollment it is important to observe whether or not the curve fitted is the curve which ought to be fitted. The test of goodness of fit is desirable in order to see if the enrollment trend for a given period is a representative one. The task of predicting future enrollment for a long period often becomes increasingly difficult because of a good many independent variables, not known at present, which may affect them in the future. In such a case, the predictions become cumulatively less reliable as time goes farther and farther. Hence, when an enrollment growth curve has been successfully fitted to a long series of data, it is preferable to compute the estimated enrollment for a short period of years as a tentative forecast. Such estimates may be proved to be close enough for most practical purposes, but special circumstances may modify their accuracy.

It appears to me that the problem of forecasting in statistics seems to be pretty much the same as the problem of being in philosophy. All of the famous philosophers had failed to solve the problem of being: Plato tried to solve it and failed; Aristotle tried and failed. In the field of sciences, many statisticians and mathematicians had tried to solve the problem of future and failed. For instance, Professor Verhulst himself had failed in fitting the population of Russia from 1796 to 1827 to a curve and abandoned the investigation of the curve of population. More recently Dr. T. C. Hungate of the Teachers College of Columbia University failed to prophesy the total college enrollment of 1950 in his book, *Financing the Future of Higher Education*, which was published in 1946. Professor R. S. Vaile of the University of Minnesota failed to foretell the 1948 enrollment in his article, "Enrollment After the War," published in 1944. The failure of forecasts from a curve to agree with the actual enrollment may be due either to the choice of the wrong type of curve or to our inability to control all factors in operation. Statisticians have lost their reputations by forecasting an uncertain event, such as predicting future college enrollment. Whenever this happens, there has been often a tendency for people to draw a hasty conclusion that statistics is no good. In view of this widespread criticism on statistics, it is deemed advisable to point out that there is no necessary relation between the goodness of fit of a mathematical curve and its reliability for forecasting. A curve may fit the enrollment data for the past fifty years with a high degree of accuracy, and yet fail to predict the future enrollment for the for the next year or two.

The forecasting of future enrollment is not simply a statistical problem and therefore perhaps never can be settled by statistics alone. The statistical estimate gained from the study of past enrollment records must be supplemented by other knowledge which may be quite nonstatistical in nature. The factors actually represented by the enrollment trend are very complex and difficult to analyze. In fact, the future college enrollment involves the operations of a rather large number of factors. For example, the number of registered students in a given college may be large or small depending upon changes in international situations and in the economic condition, veteran enrollments, birth rates, high school enrollments, unusual migration, provisions for educational benefits, policies concerning student deferment and many other numerous factors. Some of these factors may operate continuously for a number of years and some may drop out temporarily or permanently while some unknown new ones may make their appearance in the future. It appears, therefore, impossible for the statisticians to control all new factors unknown to them and consequently every statistician may make errors in probabilities at times.

In a certain sense an enrollment trend may be considered as an established habit; it is likely to continue, but other factors may change it. The element of personal judgment based upon an understanding of the factors affecting the enrollment is always invaluable. Each college should study most realistically its enrollment outlook in the light of past, present and future conditions which are likely to affect its enrollment record. If and when some estimates of the future enrollment must be made and no other device can meet the need as effectively as the fitting a trend, then projecting future enrollment under normal growth conditions by curve-fitting procedure as suggested in this paper may be exceedingly useful for the practical purposes. Finally, it should be realized that the object of forecasting enrollment is not to determine a curve that will tell exactly what must happen in the future years, but it is to make a careful analysis, based on an appropriate statistical technique, which will enable the administrator to take into account future enrollments to a greater extent than he could without them.

REFERENCES

1. Gompertz, B. "On the Nature of the Function Expressive of Human Mortality and on a New Mode of Determining the Value of Life Contingencies," Philosophical Transactions of the Royal Society of London (1825), pp. 513-85.
2. Hotelling, H. "Differential Equations Subject to Error and Population Estimates," Journal of the American Statistical Association, XXII (1927), pp. 283-314.
3. Pearl, Raymond, and Reed, L. J. "On the Mathematical Theory of Population Growth," Metron, III (1923), pp. 6-19.
4. Schultz, Henry. "The Standard Error of a Forecast from a Curve," Journal of the American Statistical Association, XXV (1930), pp. 139-85.
5. Verhulst, P. F. "Notice Sur la loi que la Population Suit dans son Accroissement," correspondance mathematique et Physique Publiee par a quetelet, Tome X (1838), pp. 113-21.

AN APPLICATION OF MULTIPLE REGRESSION ANALYSIS IN DETERMINING THE RELATIVE CONTRIBUTION OF CERTAIN COMPONENTS OF READING ABILITY IN GRADE AND HIGH SCHOOL ACHIEVEMENT

LAVERN L. KRANTZ
Ohio University
Athens, Ohio

THE PURPOSE of the present article is to show how multiple regression analysis can be applied to reveal the relationships between reading and basic skills as measured in the seventh grade (the independent variables) and success in certain phases of the academic program at the high school level (the dependent variables). For complete results of this investigation the reader is referred to a doctoral dissertation completed in 1954.[1] The main purpose of the study was to determine the significant partial regression coefficients and determine their share of the accountable variance.

The study was designed to find the relationships among measured areas of seventh-grade reading abilities and study skills and the content areas of the high school. The students of the Austin Public Schools constitute the population of this study. It is to this population that the conclusions are drawn. The samples consisted of the pupils from two seventh-grade classes. Data were obtained from the permanent records of the pupils. The first seventh-grade class tested in 1947 was tested again in 1952 when in the eleventh grade. The second seventh-grade was tested in 1949 and again in 1952 when in the ninth grade. The samples of 471 pupils included in the study, consisted of 215 pupils in the first group (7-11), and 256 pupils in the second group (7-9).

Tests used in the study were the California Intelligence Tests, (non-language section), 1946 Revision Grades 9 to Adult; the Iowa Tests of Basic Skills, Form Z and Form R; and the Iowa Tests of Educational Development, Form Y-2.

The reading abilities and work study skills measured by the Iowa Tests of Basic Skills in the Seventh Grade and their notation (Independent Variables) were as follows:

X_1	Reading Comprehension
X_2	Reading Vocabulary
X_3	Map Reading
X_4	Use of References
X_5	Use of Index
X_6	Use of Dictionary
X_7	Reading Graphs, Tables, and Charts
T_1	Total Work Study Skills
X_8	Usage (language)
X_9	Spelling
T_2	Total Language Skills
X_{10}	Fundamental Knowledge in Arithmetic
X_{11}	Fundamental Operations
X_{12}	Problems
T_3	Total Arithmetic

The measures in content areas at high school level, as obtained from scores on the Iowa Tests of Educational Development (Dependent Variables), were as follows:

Y_1	Understanding of Basic Social Concepts
Y_2	Background in the Natural Sciences
Y_3	Correctness and Appropriateness of Expression
Y_4	Ability to do Quantitative Thinking
Y_5	Ability to Interpret Reading Materials in the Social Studies
Y_6	Ability to Interpret Reading Materials in the Natural Sciences
Y_7	Ability to Interpret Literary Materials
Y_8	General Vocabulary
Y_9	Total of Above
Y_{10}	Use of Sources of Information

The assumption underlying the validity of the use of multiple regression analysis is that the dependent variable is normally distributed and that the regression of the dependent variable on the linear combination of independent variates is linear. These assumptions were tested empirically by plotting the dependent variables on normal probability paper indicating normality by a straight line relationship.

---

[1] LaVern L. Krantz. The Relationship of Reading Abilities and Basic Skills of the Elementary School to Success in the Interpretation of the Content Materials in the High School. Unpublished Ph.D. dissertation, University of Minnesota, 1954.

In order to handle the large amount of data efficiently and accurately, it was transferred to the Hollerith card which permitted the machine calculation of the squares, sums of squares, and cross products essential to the multiple regression analysis.

The first step in the analysis of the data was to determine the zero order correlations and test their significance at the one percent level of probability. The null hypothesis that the correlation coefficients are not significantly different than zero was tested.

Two main areas were investigated in each grade using the independent variates and dependent variables listed. The first multiple regression analysis was set up using X_1, X_2, X_3, T_1, T_2, T_3, and their relationship with each of the dependent variables Y_1 to Y_{10} determined. The second analysis was made using X_4, X_5, X_6, X_7, and X_8 in a similar manner with the dependent variables. It is important to point out at this time that no relationship between the two can be discussed because each one is determined from a separate matrix. The first analysis was used to obtain total relationships and the second one to obtain special relationships that may exist between each of the work study skills and the dependent variables. The second main area was a repetition of the first performed analysis for grade eleven conducted in a similar fashion.

An Example of the Application of Multiple Regression Analysis in Calculating the Reading and Study Skills Related to Y_2

The regression equation is in terms of the population values or parameters. The best estimates of these parameters are obtained from the sample values. The partial regression coefficients represent the weighted contribution that each independent variate makes to the total estimate of the dependent variable. The equation will predict the average change in the dependent variable for a unit change in anyone of the independent variates. The predictive factors involved in making a best estimate of future success in the problem under investigation is of considerable interest to those concerned with the teaching process.

The correlation between the criterion, or value we are trying to predict, and the prediction obtained by using the best weightings for the partial regression coefficients is called the multiple correlation coefficient. The multiple correlation coefficient gives an over-all evaluation of the goodness of fit of the multiple regression model. It is an indication of the degree of linear relationship between the actual scores and the best predicted scores. A test of statistical significance was applied to all multiple correlation coefficients at the one percent level of probability. If the coefficients were found not significantly different than zero the relationship could be attributed to the operation of chance factors. The square of multiple correlation coefficient gives an accurate measurement of the total amount of variance of the dependent variable mathematically accounted for by the linear combination of the independent variables. The direct contribution of each variate is obtained from the square of the standard partial regression coefficient for each variate. A major portion of the interpretation of results is based on this fact.

The next logical step after finding the reading abilities and skills apparently related to a particular high school content area was to test the significance of each relationship. When the above relationships cannot be explained on the basis of sampling errors alone they are said to be statistically significant and are, as a consequence, retained in the regression equation. An arbitrary probability level was established at the five percent level for the retention of regression coefficients used in determining the assignable variance.

The problem investigated required for its solution, not only the establishment of the existence of a relationship between reading abilities and skills and the subject areas in high school but the extent of that relationship. While two or more abilities may be contributing to the total variance of the dependent variable in high school it must be ascertained if they contribute similar amounts, different amounts, and to what degree. All partial regression coefficients stable at the five percent level, or one percent level of statistical significance as dictated by the problem solution, were tested further to determine if there was a difference in the amount of variance accounted for between them. If any differences can be accounted for on the basis of random errors of sampling, it is not possible to make any definite statements regarding the unique amount of contribution made by each one. If the difference between them cannot be accounted for by sampling error, it is inferred that one may make a greater contribution than the other.

The method of determining the amount of contribution of each variate to the total variance is shown in the sample analysis. The partial regression coefficients are changed to standard partial regression coefficients, squared, and the percent of assignable variance determined. The calculations have been completed in the example to illustrate the procedures used.

Steps commonly followed in the solution of a multiple regression problem in a stable solution at the five percent level of statistical significance for the regression coefficients are not shown and may be found in other sources.[2] The computations below carry the solution on using the partial regression coefficients found to be significant. In the example used it was found that the partial regression coefficients for independent variables T_1 (Work Study Skills), T_3 (Total Arithmetic Skills), and X_3 (Vocabulary) were significantly related to Y_2 (Background in Natural Science) and accounted for 51 percent of the variance.

Our first requirement is that of converting the partial regression coefficients to standard partial regression coefficients.[3] This is required in order to render the measurements independent of scale. The process of transformation is indicated below where $\beta(i = 1, 2, 3)$ represents the respective standard partial regression coefficients.

$$\beta_1 = .1301 \sqrt{\frac{T_1^2}{Y_2^2}} = .1301 \sqrt{\frac{65714.7343}{7203.2773}} = .3929$$

$$\beta_2 = .1071 \sqrt{\frac{T_3^2}{Y_2^2}} = .1071 \sqrt{\frac{24153.4960}{7203.2773}} = .1963$$

$$\beta_3 = .0522 \sqrt{\frac{X_3^2}{Y_2^2}} = .0522 \sqrt{\frac{93114.3398}{7203.2773}} = .1877$$

We next square each of the standard partial regression coefficients and calculate the proportion that each square constitutes of the total variance, i.e., R^2. (See table below.) The change to a percentage basis is indicated in the third column.

Summary and Conclusions

Multiple regression analysis was applied to determine the relationships among seventh-grade reading abilities and study skills and achievement in the content areas of the high school.

After running the usual analysis and finding the partial regression coefficients stable at the five percent level of statistical significance, further treatment was applied to determine the extent of the relationships. This was accomplished by testing the significance of the difference between partial regression coefficients to ascertain if a real difference in their respective contributions existed or if it could be accounted for by random errors of sampling. An example was shown illustrating the method of calculating the percentage of the total variance accounted for by each component. The above procedures used in treating the data were of basic importance in the interpretation of the results.

	SP^2	Proportion	Percent
T_1 : Work Study Skills	.1543	.6767	67.6
T_3 : Total Arithmetic Skills	.0385	.1688	16.8
X_3 : Vocabulary	.0352	.1543	15.4
	.2280	.9998	99.8

---

[2] Palmer O. Johnson. Statistical Methods in Research (New York: Prentice-Hall, Inc., 1949), 377 pp.

[3] Palmer O. Johnson. "The Differential Functions of Examinations," Journal of Educational Research, XXX (October 1936), pp. 93-103.

A REDUCED FORMULA FOR POINT-BISERIAL COEFFICIENT OF CORRELATION

WILLIAM O. BUSCHMAN
Portland State Extension Center
Portland, Oregon

TWO OF THE most common types of correlation coefficients for the purpose of item analysis are biserial *r* and point-biserial *r*. The first of these, biserial *r*, assumes that the item is represented by a dichotomization of a continuous and normally distributed variable and that the criterion is a continuous but not necessarily normal variable. Whether or not these assumptions can be justified is often questionable. Richardson and Stalnaker (5:463-64) indicate that the biserial *r* is not a suitable measure for item analysis where the responses are dichotomized because it assumes: (1) that the distribution of the dichotomized variable is normal, and that (2) the dichotomized variable is continuous. They indicate that unless these conditions can be established and be shown to fit the case being studied, that use of the biserial *r* is not warranted. In such cases the point-biserial *r* is more suitable.

The laborious computations which are required for most methods of item analysis have been a deterrent to their use. In order to reduce the labor of computing the biserial coefficient, Flanagan devised a procedure using only the upper and lower portions of the group (3: 674-80). Dunlap (2:51-58), Chapanis(1:297-304) and others have proposed methods for shortening the amount of computation in determining the biserial index. Such shortened methods are preferable if they yield equally satisfactory results, and in some cases may be preferred because of the saving in time and effort even though a certain degree of accuracy may be sacrificed. A reduced formula for the point biserial coefficient is developed and shown below.

The point-biserial coefficient of correlation is given by the following formula (4:353):

$$r_{rpb} = \frac{M_p - M_f}{\sigma_T} \sqrt{pq} \qquad (1)$$

in which:

M_p = the mean score of the individuals passing the item,

M_f = the mean score of individuals failing the item,

σ_T = the standard deviation of the total distribution of scores,

p = the proportion of individuals passing the item,

q = the proportion of individuals failing the item.

For purposes of comparing the validity of items on a single test and not for comparison with items on other tests, the reduction to standard form by division by σ_T is unnecessary and may be omitted. This gives what will be termed the reduced point-biserial coefficient of correlation. The formula for this may then be written:

$$r_{rpb} = (M_p - M_f)\sqrt{pq} \qquad (2)$$

The difference $M_t - M_f$ is as adequate as $M_p - M_f$ (1:303). Thus the formula can be expressed in the alternative form:

$$r_{rpb} = (M_t - M_f)\sqrt{pq} \qquad (3)$$

in which M_t is the mean of the total distribution of scores.

If deviations are taken from the same step in each case, then the means M_t and M_f may be replaced by the deviations of the true means from the guessed means as the differences will be cancelled in the subtraction $M_t - M_f$. The formula can then be expressed:

$$r_{rpb} = \left(\frac{\Sigma f_t d}{N_t} - \frac{\Sigma f_f d}{N_f}\right)\sqrt{pq} \qquad (4)$$

in which N_t represents the total number of scores and N_f the number failing the item. This is a more usable form than either (2) or (3) above. Further simplification of the computations are possible by taking the deviations (d) in terms of class intervals since the class interval is constant throughout any problem. If the zero deviation is taken at the lowest interval in which any frequencies occur, the negative deviations may be avoided.

The reduced formula will give values which are just as accurate and adequate as the formulas from which they are derived. Though results can be applied only to the problem on which the values are computed, this is the way in which this coefficient is generally used. If comparisons with other sets of data are sought, the data obtained by the reduced formula can be converted to the conventional formulas, but only in rare cases would this furnish useful information.

The development of the reduced form of the formula for the point-biserial coefficient of correlation (4) will make it easier to use and apply this coefficient to those situations in which it is the most suitable of the several measures which are used for item analysis.

REFERENCES

1. Chapanis, Alphonse. "Notes on the Rapid Calculation of Item Validities." Journal of Educational Psychology 32:297-304. 1941.
2. Dunlap, J. W. "Note on the Computation of Bi-serial Correlation in Item Evaluation." Psychometrika 1:51-58. 1936.
3. Flanagan, John C. "General Considerations in the Selection of Test Items and a Short Method of Estimating the Product Moment Coefficient From the Data at the Tails of the Distributions." Journal of Educational Psychology 30:674-680. 1939.
4. Garrett, Henry E. Statistics in Psychology and Education. 3rd. ed. New York, Longmans, Green and Co., 1947. 487 p.
5. Richardson, M. W. and J. M. Stalnaker. "A Note on the Use of Bi-serial r in Test Research." Journal of General Psychology 8:463-465. 1941.

Journal of Experimental Education

Volume XXIII June, 1955 Number 4

PREDICTION OF THE PROBABILITY OF GRADUATION OF ENGINEERING TRANSFER STUDENTS

J. STANLEY AHMANN
Cornell University

THERE IS little doubt in the minds of many that the educator of today is deeply interested in the prediction of the future. Even a casual examination of the literature readily demonstrates the willingness of investigators in education to devote considerable time and effort in attempts to determine their ability to predict the educational scene of tomorrow from known data of today. Specifically, repeated studies have been made to determine estimates of the future academic achievement of students, those estimates being based upon such present measures as scholastic aptitude, mechanical aptitude, interest, and previously demonstrated scholastic ability.

It is evident that the results of these efforts to predict, most of which have been based upon principles of correlation and linear regression, have assisted the instructors and counselors of students. This is true for the most part in engineering education as well as in other areas. Needless to say, however, the degree of assistance has varied greatly. For example, predictive studies concerning certain groups of students are comparatively few in number. One such group is the transfer student group, especially the transfer students in engineering.

Although education research and editorial comment reported in the higher education journals tend to exclude the question of the position of transfer students in colleges and universities, the importance of this group of students is steadily increasing in the opinions of many. In engineering, for example, instructors and administrators alike are becoming more and more aware of the impact on engineering colleges made by the undergraduate students who first began their college studies at institutions of higher learning other than the one from which they ultimately hoped to graduate. The academic and social adjustments to be made by such students who transfer from one institution to another are often genuine problems for the instructors and counselors of both institutions concerned.

The Division of Engineering at the Iowa State College enrolls sizeable numbers of transfer students each academic year. Estimates have been made that as many as 40 percent of the entering engineering students at the beginning of a fall quarter had received college credit from other institutions of higher learning. The percent of the engineering graduating class who were originally transfer students has been as large as 50 percent on occasion.

In the light of these figures alone, further investigating of the collegiate academic careers of the transfer students entering engineering would seem warranted. In addition, college officials and transfer students have repeatedly inquired about the likelihood of a student's academic success in engineering after transfer. The answers to such questions have been shrewd guesses at best since few adequate tools have been developed which will reduce to administrative routine the prediction of a transfer student's academic success in the Engineering Division.

It was the purpose of this study (1), therefore, to devise an instrument to predict the probability of academic success in engineering of students who transferred from other institutions of higher learning. For the purpose of this study, academic success has been defined in terms of whether the transfer student eventually graduated in engineering at the Iowa State College. It should be noted that this dichotomous criterion of graduation-nongraduation differed from the usual numerical criterion of course marks or cumulative grade point averages used in regression studies (4). However, the definition here chosen is considered to be reasonable in light of the fact that some engineering faculties seem to be more interested in whether a student graduates in engineering than whether he attains some numerical academic average above the minimum required for graduation. Furthermore, in view of the present day shortage of trained engineers, certain industrial concerns are inclined to consider a student's ability to graduate in engineering from a specific institution as adequate demonstration of academic achievement. Lastly, such a simple definition of academic achievement appeals to the entering transfer student who some day may be counseled on the basis of these results.

Prediction of the probability of graduating in engineering was based upon various combinations of test scores compiled by the psychological testing program, the high school grade point average, a rating of the student's previous collegiate academic achievement, and the first quarter grade point average in engineering at the Iowa State College. In all, six prediction variables were considered. Four of the variables were assembled as a matter of routine and were available prior to the date on which the transfer students began their studies at the Iowa State College. They were the linguistic raw scores and quantitative raw scores on the American Council on Education Psychological Examination 1945 Edition, the high school grade point averages, and the raw scores on the United States Armed Forces Institute Test on Correctness and Effectiveness of Expression, College Level.

The fifth prediction variable which also was available prior to the beginning of the transfer student's Iowa State College studies, was a rating of each student's prior college achievement as represented by the transcript of his credits. This rating was made on a six point scale by clerks in the admissions branch of the registrar's office, and was based upon the type of courses previously completed by the student, the final marks he received, and the engineering curriculum he was entering at the Iowa State College.

The sixth prediction variable was the first quarter grade point average at the Iowa State College. Because the transfer students entered the Engineering Division with a variety of collegiate backgrounds, the first quarter grade point average by no means represented academic achievement in a uniform set of courses. The five point letter marking system used in both the high school marks and the first quarter college marks was converted to a 4, 3, 2, 1, 0 scale.

Although the prediction variables resemble closely those used in regression schemes, the criterion, the tendency to graduate in engineering, was a dichotomy rather than the usual numerical variable. Consequently the regression technique was not applicable, and the discriminant function as reported by Fisher (2) was considered. Fisher developed the discriminant function for the purpose of finding the weights to be assigned to a series of variables that would produce maximum separation between two groups, each of which is assumed to be normally distributed. In the case of the present criterion maximum separation was also desired but within a single normally distributed group which had been divided into a dichotomy.

The foregoing assumption that the tendency to graduate in engineering is a single normally distributed variable necessitated a modification of Fisher's discriminant function, such a modification having been reported in a forthcoming textbook by Wert, Neidt, and Ahmann (6). This modification yielded a prediction equation resembling greatly the regression approach, with the major exception that the prediction is made in terms of chances in 100 of graduating in engineering rather than in terms of achievement scores. The general form for the modified discriminant function of *m* prediction variables is:

$$V_O = a_1X_1 + a_2X_2 + a_3X_3 + \dots + a_mX_m + C$$

where

V_O = sigma units
a = coefficients
X = prediction variable scores
C = constant

Substitution of the appropriate values of the precition variables yields a prediction in terms of sigma units. By consulting a table of the normal curve, the sigma units are converted to chances in 100.

To determine the size of the coefficients which would yield the best possible prediction equation, data was gathered for 804 male engineering transfer students who enrolled since September, 1946, and who at a later date, permanently left the Engineering Division either by means of graduation or dropping out. Of the 804, over three-fourths were veterans of World War II. In general, most of the students had attended only one college prior to enrolling at the Iowa State College, and the college was usually located in Iowa or an adjacent state and enrolled less than 2500 students. Approximately one-third of the students first attended junior colleges before transferring whereas the remainder first matriculated at colleges other than junior colleges.

Since earlier research (3) has indicated that veteran students tend to surpass non-veteran students in academic achievement, the original group of 804 was so divided. In addition preliminary investigation of the available data revealed the possibility that, when individual differences are ignored, those transfer students who first attended junior colleges experienced greater difficulty in graduating in engineering than those who first attended colleges other than junior colleges. Although the reasons for this situation were not determined, the veteran and non-veteran transfer students were further stratified on the basis of college status. The resulting four subgroups were:

1. Veteran students from junior colleges (k = 207)

TABLE I

TENDENCY TO GRADUATE IN ENGINEERING OF 804 TRANSFER STUDENTS DIVIDED INTO FOUR SUBGROUPS

	Graduates		Nongraduates		Total	
Subgroup	No.	%	No.	%	No.	%
Veteran, Junior College	52	25.1	155	74.9	207	100.0
Non-Veteran, Junior College	8	13.1	53	86.9	61	100.0
Veteran, Other College	159	38.4	255	61.6	414	100.0
Non-Veteran, Other College	27	22.1	195	77.9	122	100.0
Total	246	30.6	558	69.4	804	100.0

TABLE II

WITHIN SUBGROUPS BISERIAL CORRELATIONS BETWEEN THE CRITERION AND ALL PREDICTION VARIABLES

Prediction Variable	Biserial Correlation
Linguistic Score on ACE Test	.187
Quantitative Score on ACE Test	.296
High School Average	.312
USAFI English Score	.281
Admissions Office Rating	.396
First Quarter Average	.655

TABLE III

MULTIPLE BISERIAL COEFFICIENTS OF CORRELATION RESULTING FROM VARIOUS DISCRIMINANT FUNCTIONS

Combinations of Prediction Variables	Multiple Biserial Coefficients
Q-score, High School Average	.380
Q-score, High School Average, USAFI Eng.	.399
Q-score, High School Average, USAFI Eng., Rating	.480
Q-score, High School Average, USAFI Eng., Rating First Quarter Average	.686

TABLE IV

CHANCES IN 100 OF GRADUATING IN ENGINEERING FROM PRE-MATRICULATION DATA

H. S. Ave.	Sub-Group*	ACE Q-Score Percentile Rank					
		1-10	11-20	41-50	51-60	81-90	91-100
Below 1.60	1	13	15	18	19	23	26
	2	6	8	10	10	13	15
	3	21	24	29	30	35	38
	4	8	10	12	13	16	19
1.60-1.79	1	14	16	20	21	25	28
	2	7	8	11	11	14	16
	3	23	26	30	32	37	40
	4	9	11	14	14	18	20

2.40-2.59	1	19	22	26	28	32	36
	2	10	12	15	16	20	22
	3	30	33	38	40	45	48
	4	13	15	19	20	24	27
2.60-2.79	1	21	24	28	29	34	38
	2	11	13	16	17	21	24
	3	32	35	40	42	47	50
	4	14	17	20	21	26	28

3.60-3.79	1	29	33	38	39	45	48
	2	17	20	24	25	30	33
	3	42	45	51	52	58	61
	4	21	24	29	30	35	38
3.80-4.00	1	31	35	40	41	47	50
	2	19	22	26	27	31	35
	3	44	48	53	54	60	63
	4	23	26	31	32	37	40

* 1 = Veteran, junior college
2 = Non-veteran, junior college
3 = Veteran, other college
4 = Non-veteran, other college

TABLE V

CHANCES IN 100 OF GRADUATING IN ENGINEERING FROM FIRST QUARTER GRADE POINT AVERAGES

1st Quarter Average	Subgroup	Chances in 100
0.00-0.19	1	11
	2	7
	3	16
	4	9
0.20-0.39	1	12
	2	8
	3	18
	4	10

1.80-1.99	1	32
	2	25
	3	42
	4	28
2.00-2.19	1	35
	2	28
	3	46
	4	31

3.60-3.79	1	63
	2	54
	3	72
	4	58
3.80-4.00	1	66
	2	58
	3	75
	4	62

2. Non-veteran students from junior colleges ($k = 61$)
3. Veteran students from other colleges ($k = 414$)
4. Non-veteran students from other colleges ($k = 122$)

Examination of Table I readily reveals the fact that the four subgroups tended to respond to the criterion differently. The stratification was therefore maintained throughout the statistical analysis.

Since the sample had been stratified into four subgroups, to determine the relationship between the tendency to graduate in engineering and each of the prediction variables, within subgroups biserial coefficients of correlation were computed and are listed in Table II. All relationships were found to be significantly different from zero.

Modified discriminant functions using various combinations of the prediction variables were computed for within the subgroups and yielded multiple biserial coefficients of correlation indicative of the predictive ability of each scheme. It was immediately discovered that the linguistic score could be dropped from any prediction combination without incurring an appreciable loss in ability to predict. In Table III are shown the various combinations of prediction variables and the resulting multiple biserial coefficients of correlation. The correlations are indicative of the ability to predict for any of the four designated subgroups since, in the case of any of the foregoing discriminant functions, the coefficients ($a_1, a_2, \ldots, a_m$) to be multiplied by the appropriate prediction variable scores were the same for each subgroups. However, the constant (C) changed with each subgroup so that the predictions vary from one subgroup to another.

To facilitate the application of the discriminant function technique for such persons as counselors of students, tables of chances in 100 of graduating were computed by repeatedly solving a discriminant function. By means of a single table or a series of related tables a counselor can rapidly determine a transfer student's probability of graduating in engineering by entering the table or tables with the student's prediction variable scores. If the prediction were to be made on the basis of two prediction variables, a single table such as Table IV would be sufficient. For three variables, two tables are necessary. The number of probability tables to be consulted is one less than the number of variables to be used in the prediction when more than one prediction variable is involved.

If, prior to the date that a transfer student began his studies at the Iowa State College, it were desirable to predict his probability of graduating in engineering on the basis of his quantitative score and high school grade point average, a counselor could easily do so by means of a table such as that partially shown in Table IV. Because the quantitative scores are usually available as percentile ranks, the table was constructed accordingly.

The probabilities were computed for each of the four subgroups, ranging from 6 chances in 100 for the least qualified students to 63 chances in 100 for the best qualified students, and varying appreciably from subgroup to subgroup. In using the table the counselor would classify a transfer student in the appropriate subgroup, read up and across in accordance with the student's scores, and note his chances in 100 of graduating in engineering. This information based upon the academic success of an earlier group of transfer students with similar qualifications, certainly would provide the incoming transfer student with a more realistic basis on which to evaluate his position.

If a counselor were willing to wait until the completion of a transfer student's first quarter of studies at the Iowa State College before predicting the student's probability of survival, a table such as the one shown in part in Table V could be used. Table V resembles closely Table IV except that the prediction is now being made exclusively on the basis of the first quarter grade point average. On the basis of this variable, the probabilities of graduating ranged from 7 chances in 100 to 75 chances in 100. Again the probabilities differed among the subgroups.

The application of the modified discriminant function technique to the foregoing problem is but one of many applications. The prediction of the probability of attending college or of passing a specific course (5) are examples of prediction problems in which the investigator may wish to assume that the dichtomous criterion is in reality a single normally distributed variable. If, as in the case of the prediction of the probability of graduation, the ability to predict is found to be sizable, certainly the application of the technique to these and similar problems provides the instructor, the administrator, and the counselor with still another valuable tool.

REFERENCES

1. Ahmann, J. Stanley. Prediction of Achievement of Iowa State College Engineering Students Having Transferred from Other Institutions. Unpublished Ph. D. Thesis, Iowa State College, 1951.
2. Fisher, R. A. "The Use of Multiple Measurements in Taxonomic Problems," Annals of Eugenics, VII (1936), pp. 179-88.
3. Gowan, A. M. Unique Characteristics of

Freshman Veterans at the Iowa State College with Administrative Implications. Unpublished Ph. D. Thesis, Iowa State College, 1947.

4. Moore, J. E. "A Decade of Attempts to Predict Scholastic Success in Engineering Schools," Occupations, XXVIII (November 1949), pp. 92-6.

5. Sprain, Wilbur. Forecasting Probability of First Quarter Mortality in General Chemistry at Iowa State College. Unpublished M. S. Thesis, Iowa State College, 1951.

6. Wert, James E., and others. Statistical Methods in Education and Psychology (New York: Appleton-Century-Crofts Co., 1953).